

Identifying the effect of persuasion

Sung Jae Jun
Sokbae Lee

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP24/22

IDENTIFYING THE EFFECT OF PERSUASION*

SUNG JAE JUN[†]

SOKBAE LEE[‡]

PENN STATE UNIV.

COLUMBIA UNIV. AND IFS

November 30, 2022

Abstract. This paper examines a commonly used measure of persuasion whose precise interpretation has been obscure in the literature. By using the potential outcome framework, we define the causal persuasion rate by a proper conditional probability of taking the action of interest with a persuasive message conditional on not taking the action without the message. We then formally study identification under empirically relevant data scenarios and show that the commonly adopted measure generally does not estimate, but often overstates, the causal rate of persuasion. We discuss several new parameters of interest and provide practical methods for causal inference.

Key Words: Communication, Media, Persuasion, Partial Identification, Treatment Effects

JEL Classification Codes: C21, D72, L82

*We would like to thank the editor, four anonymous referees, Eric Auerbach, Stefano DellaVigna, Leonard Goff, Marc Henry, Keisuke Hirano, Joel Horowitz, Charles Manski, Joris Pinkse, Imran Rasul, Roman Rivera, Myunghyun Song, David Yang, and seminar participants at Northwestern University, Rutgers University, Seoul National University and Vanderbilt University for helpful comments. This work was supported in part by the European Research Council (ERC-2014-CoG-646917-ROMIA) and by the UK Economic and Social Research Council (ESRC) through research grant (ES/P008909/1) to the Centre for Microdata Methods and Practice. This paper includes applications based on previously published articles, and we would like to thank the authors of these papers for making their data sets and replication files available via journal archives or personal web pages.

[†]Department of Economics, Pennsylvania State University, 619 Kern Graduate Building, University Park, PA 16802. Email: suj14@psu.edu

[‡]Department of Economics, Columbia University, 1022 International Affairs Building, 420 West 118th Street, New York, NY 10027; Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK. Email: sl3841@columbia.edu

1. INTRODUCTION

How effectively one can persuade one's audience has been of interest to ancient Greek philosophers in the Lyceum of Athens,¹ early-modern English preachers in St Paul's Cathedral,² and contemporary American news producers at Fox News in New York City.³ Recently, economists have been endeavoring to build theoretical models of persuasion (e.g., Kamenica and Gentzkow, 2011; Che, Dessein, and Kartik, 2013; Gentzkow and Kamenica, 2017; Prat, 2018; Bergemann and Morris, 2019) and to quantify empirically the extent to which persuasive efforts affect the behavior of consumers, voters, donors, and investors (see DellaVigna and Gentzkow, 2010, for a survey of the recent literature).

Since DellaVigna and Kaplan (2007, DK hereafter) proposed a measure of persuasion, it has been used and modified by many authors (e.g., Enikolopov, Petrova, and Zhuravskaya, 2011; Gentzkow, Shapiro, and Sinkinson, 2011; DellaVigna, Enikolopov, Mironova, Petrova, and Zhuravskaya, 2014; Bassi and Rasul, 2017; Martin and Yurukoglu, 2017; Chen and Yang, 2019) to quantify the persuasive effects of informational treatment. However, its precise interpretation has been obscure because of a lack of formal identification analysis. In fact, we show that the commonly used measure of persuasion does not estimate the causal rate of persuasion on any subpopulation in general. Therefore, it is misleading to call DK's measure the persuasion rate, although it is a common practice in the literature; instead, we will reserve the term of the persuasion rate for the population parameter that is properly defined by a conditional probability in the potential outcome framework.

The flaw of DK's measure arises from failing to distinguish a local average treatment effect (LATE; see Imbens and Angrist, 1994) from the the average treatment effect (ATE), where the difference between the two can be substantial for a heterogeneous population; specifically, DK's measure rescales a LATE with a factor that is relevant only for the ATE. For instance, in DK's example, even if Fox News has a high persuasive effect among those who will watch the channel if and only if it is available through a local cable package, it may not be persuasive at all when other people such as Democrats or Democrat-leaning independents are all included.

Focusing on the case of binary outcomes, we analyze the problem of measuring persuasive effects of informational treatment through the lens of the potential outcome framework. Specifically, we define the persuasion rate by a proper conditional probability of the agent taking an action of interest with a persuasive message given that the agent does not take the action without the conveyed message. We then formally study identification under a few empirically relevant scenarios of data availability. Our analysis will articulate what DK's measure of persuasion estimates, why it is misleading, and how we can fix the problem.

¹See Rapp (2010) for three technical means of persuasion in Aristotle's *Rhetoric*.

²See Kirby (2008) for historic details of the public persuasion at Paul's Cross, the open-air pulpit in St Paul's Cathedral in the 16th century.

³DellaVigna and Kaplan (2007) and Martin and Yurukoglu (2017) measure the persuasive effects of slanted news using data on Fox News.

While the persuasion rate is concerned with the entire population (and hence related with the ATE), we also consider a local persuasion rate that focuses on the group of compliers. Our identification analysis shows that DK's measure estimates neither a local nor an average persuasion rate.

Our identification analyses are based on a few empirically relevant scenarios of data availability. We do this because the problem of data availability is particularly important in the context of measuring persuasive effects. For instance, individual-level partisan vote outcome data rarely exist due to confidentiality issues. Thus, the outcome variable is frequently measured only at an aggregate level. Also, it is not always the case to observe an actual exposure to a persuasive message in the same data set along with the outcome and instrument. Indeed, DK's analysis uses a micro dataset for the treatment and instrument and a separate aggregate dataset for the outcome and instrument. In order to address these challenges, we consider three different scenarios of data availability explicitly: given the instrument, (i) the outcome and treatment are jointly observed, (ii) they are observed separately, or (iii) the treatment is not observed at all. We obtain the sharp bounds on the persuasion rate for the entire population as well as other subpopulations of potential interest. Therefore, our work builds on the econometrics literature on partial identification (e.g., Manski, 2003, 2007; Tamer, 2010) as well as the literature on program evaluation (see Heckman and Vytlacil, 2007a; Imbens and Wooldridge, 2009, for surveys of the literature).

The main findings of this paper can be summarized as follows. If there is no heterogeneity in the population, then DK's measure of persuasion estimates the rate of persuasion, provided that a simple monotonicity assumption is imposed; however, this case is an exception rather than the rule. Indeed, the rate of persuasion is only partially identified as an interval in general, where the sharp lower bound generally corresponds to DK's measure multiplied by the relative size of the complier group regardless of the specific data scenario. Therefore, DK's measure is strictly larger than the lower bound whenever there is partial compliance. It is also remarkable that the data scenarios matter only for the sharp upper bound. Therefore, the value of observing the treatment and outcome jointly only lies in obtaining a potentially more informative upper bound on the persuasion rate. We also investigate identification of the local persuasion rate (i.e., the persuasion rate for the group of compliers) under the same three data scenarios. It is point-identified under the most favorable data scenario, but only partially identified under the other scenarios; even in the case of point identification, DK's measure generally differs from the local persuasion rate. If a continuous instrument is available, then we can target a *marginal persuasion rate* that is akin to the marginal treatment effect (e.g., Heckman and Vytlacil, 2005). Therefore, having a continuous instrument opens up the possibility of point identification of the persuasion rate for a policy-relevant population if the instrument is sufficiently rich.

In order to illustrate our findings, we discuss two empirical examples in the main text, while we provide a few more in the online appendices. First, we revisit Chen and Yang (2019), where the interest is in

the effect of Chinese students having access to uncensored media on behaviors, beliefs and attitudes; we use the same variables and setup as in the original paper for this exercise. Second, we analyze the voting behavior and newspaper readership using the data from Gerber, Karlan, and Bergan (2009). Overall, we show that DK’s measure of persuasion tends to overstate the persuasive effects while masking underlying heterogeneity.

The remainder of the paper is organized as follows. In Section 2, we recall some backgrounds and specifics of DK’s measure of persuasion. Here, we properly define the persuasion rate at the population level by using the potential outcome framework. In Section 3, we discuss identification of the persuasion rate. In Section 4, we study the local and marginal versions of the persuasion rate. In Section 5, we discuss our recommendations on what to do in practice, including practical inferential issues,⁴ and clarify the difference between a population version of DK’s measure of persuasion and the persuasion rate. In Section 6, we provide two empirical illustrations. We conclude in Section 7. The online appendices include additional results and examples, including an extension to non-binary outcomes, a detailed discussion about methods for inference, and all of the proofs.

2. BACKGROUND

It is helpful to recall DK as a prototypical example, where they study the effect of an exposure to Fox News on the probability of voting for a Republican presidential candidate. Here, the informational treatment of interest is the viewership of the Fox News channel, and the persuasion rate of the media can be thought of as the proportion of the Fox News viewers who voted for a Republican candidate among those who would not have done so if they had not watched Fox News. It should be noted that the agents’ decisions about whether to watch Fox News or not may be correlated with their political orientation. In order to address the endogeneity issue, DK use Fox News availability via local cable in the year of 2000 as an instrumental variable.

In efforts to measure the persuasion rate as explained above, DK and DellaVigna and Gentzkow (2010) propose the (infeasible) estimand f defined as follows: for a binary outcome and by using DK’s notation

$$f = \frac{y_{\mathbb{T}} - y_{\mathbb{C}}}{e_{\mathbb{T}} - e_{\mathbb{C}}} \cdot \frac{1}{1 - y_0}, \quad (1)$$

where \mathbb{T} and \mathbb{C} represent an instrument assignment status such as having Fox News available via local cable or not.⁵ Here, for $j \in \{\mathbb{T}, \mathbb{C}\}$, y_j is the share of group j taking the action of interest (e.g., voting for a Republican candidate), and e_j is the share of group j exposed to a persuasive message. Further, y_0 is the share of those who would take the action of interest without listening to the persuasive message. So, f

⁴Stata commands for estimation and inference based on this paper’s identification results are publicly available at <https://github.com/persuasio>. Alternatively, they can be installed from within Stata by typing “ssc install persuasio”.

⁵Therefore, \mathbb{T} and \mathbb{C} , which appear to denote treatment and control groups, should be understood as the status of the intent to treat, not the actual treatment status.

is a rescaled version of the usual Wald statistic that estimates the LATE, where rescaling is apparently to obtain a “rate” that focuses on those who are to be persuaded. As DK noted, y_0 involves a counterfactual that is often unobserved. For this reason, f is generally not a feasible estimand, and DK propose using y_C in place of y_0 as an approximation, which yields a feasible estimand, say \tilde{f} . We will refer to f or its feasible approximation \tilde{f} by DK’s measure of the persuasion rate.

Without a formal justification, it is common in the literature to interpret f as a conditional probability. For instance, DK explain, on page 1218 of their paper, “*The key parameter is f , the fraction of the audience that is convinced by Fox News to vote Republican.*” Similar interpretations are prevalent in the literature, as the following quotations demonstrate:

DellaVigna and Gentzkow (2010, p.645): *Whenever possible, we report the results in terms of the persuasion rate (DellaVigna & Kaplan 2007), which estimates the percentage of receivers that change the behavior among those that receive a message and are not already persuaded.*

Gentzkow, Shapiro, and Sinkinson (2011, p.3003): *We can also translate our estimates into a “persuasion rate” (DellaVigna and Kaplan 2007) – the number of eligible voters who changed their voting behavior as a result of the introduction of the newspaper, as a fraction of all those who could have changed their behavior.*

DellaVigna, Enikolopov, Mironova, Petrova, and Zhuravskaya (2014, p.125): *The persuasion rate is the fraction of the audience of a media outlet who are convinced to change their behavior (in this case, their vote) as a result of being exposed to this media outlet.*

Martin and Yurukoglu (2017, p.2590): *Finally, we computed estimates of DellaVigna and Kaplan’s (2007) concept of persuasion rates: the success rate of the channels at converting votes from one party to the other. The numerator in the persuasion rate here is the number of, for example, the Fox News Channel (FNC) viewers who are initially Democrats but by the end of an election cycle change to supporting the Republican party. The denominator is the number of FNC viewers who are initially Democrats.*

Also, in their survey, **DellaVigna and Gentzkow (2010)** use \tilde{f} , i.e., a feasible approximation of f , as a key summary statistic to compare persuasive effects across different studies.

However, as we mentioned in the introduction, neither f nor \tilde{f} estimates the persuasion rate, i.e., the intended conditional probability, on any subpopulation in general. Therefore, it is misleading to label f or \tilde{f} as a persuasion rate, and it is generally invalid to make comparisons across different studies. Indeed, f or \tilde{f} may not even be a proper conditional probability in a heterogeneous population: e.g., the approximation \tilde{f} can even be larger than one. We will articulate under what assumptions \tilde{f} turns out to be a proper conditional probability, and how we should interpret it in its relationship with the persuasion rate.

In order to facilitate our discussion, we start with formally defining the persuasion rate at the population level by using the potential outcome framework. Let T_i denote the binary indicator that equals 1 if individual i is exposed to persuasive information such as Fox News. Let $Y_i(t)$ be a binary indicator, which shows agent i 's potential action when T_i is set to $t \in \{0, 1\}$. For example, $Y_i(1)$ equals 1 if individual i votes for a Republican candidate after watching Fox News. The econometrician never observes both $Y_i(0)$ and $Y_i(1)$ but can only observe either of the two: that is, $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$.⁶ Then, the fraction of the people who take the action of interest with an exposure to a persuasive message, among those who would not without it, is given by

$$\theta_{\text{pr}} := \mathbb{P}\{Y_i(1) = 1 \mid Y_i(0) = 0\}, \quad (2)$$

provided that the conditional probability is well-defined: θ_{pr} is the persuasion rate at the population level. Using conditional probability is to rule out the case of “preaching to the converted”; if $Y_i(0) = 1$, then those individuals are already “persuaded” to take the action of interest even without the persuasive treatment, and therefore we do not count them in defining the persuasion rate.⁷

The common estimand \tilde{f} (or even f) does not estimate θ_{pr} in a heterogeneous population; in fact, even in a homogeneous population, \tilde{f} or f cannot be (asymptotically) equated with θ_{pr} without an extra monotonicity assumption. However, the rescaled quantity $(e_{\text{T}} - e_{\text{C}})\tilde{f}$, which is always no greater than \tilde{f} , does provide valid information about θ_{pr} in that it corresponds to the sharp lower bound of the identified interval of θ_{pr} in general. The best way to clarify all the issues is to conduct a rigorous analysis on the identification of θ_{pr} , which is our next topic. For quick takeaways, see Section 5.

3. IDENTIFICATION OF THE PERSUASION RATE

Identification of θ_{pr} is challenging for various reasons, including the fact that θ_{pr} depends on the joint distribution of the potential outcomes and that T_i can be endogenous, and it is often difficult to observe T_i and Y_i jointly. To allow for potential endogeneity, we use a binary instrument, Z_i , throughout the paper unless otherwise specified. Exogenous covariates, X_i , could be observed, but we suppress X_i from our identification analysis. In other words, we implicitly assume throughout the paper that all assumptions and results are conditional on the value of X_i . Therefore, the observed variables are Y_i, T_i, Z_i , all of which are binary throughout the main text. See Appendix D for how to deal with X_i in practice. Also, see Appendix B for an extension to the case of multinomial outcomes.

The data issue on T_i is addressed by considering three scenarios of data availability. Specifically, for the purpose of the identification analysis, we assume that for $(y, t, z) \in \{0, 1\}^3$, (i) $\mathbb{P}(Y_i = y, T_i = t \mid Z_i = z)$ is known, (ii) $\mathbb{P}(Y_i = y \mid Z_i = z)$ and $\mathbb{P}(T_i = t \mid Z_i = z)$ are known, or (iii) $\mathbb{P}(Y_i = y \mid Z_i = z)$ is all

⁶So, both Y_i and T_i are binary. In online Appendix B, we extend our results to the case where the potential outcomes are multinomial.

⁷The idea of using conditional probability to define a parameter of interest can also be found in Heckman, Smith, and Clements (1997), though their context is quite different from ours.

that is known, depending on the specific scenario of interest. For example, DK use town-level election data to estimate $\mathbb{P}(Y_i = y \mid Z_i = z)$ and micro-level media audience data to infer $\mathbb{P}(T_i = t \mid Z_i = z)$, which corresponds to Case (ii).⁸

It requires an additional assumption to address the challenge that $Y_i(1)$ and $Y_i(0)$ are never simultaneously observed while θ_{pr} depends on their joint distribution. Before we present our identification results, we discuss our key assumptions in the following subsection.

3.1. The Key Assumptions. Our first key assumption is that the persuasive message is directional, which will be important to decouple θ_{pr} by the marginals of the potential outcomes.

Assumption A (Monotonic Treatment Response). *The potential outcomes $Y_i(1)$ and $Y_i(0)$ are binary, and they satisfy $Y_i(0) \leq Y_i(1)$ with probability one.*

Assumption A is a binary version of the monotonic treatment response (MTR) assumption used in Manski (1997) and Manski and Pepper (2000).⁹ Assumption A allows $Y_i(0) = Y_i(1)$ with probability one, and therefore it does not rule out the possibility that ‘watching Fox News’ has no impact on the agent’s behavior at all. The inequality in Assumption A means that the messages Fox News delivers are *biased or directional* in favor of Republican candidates; i.e., if a voter is going to vote for a Republican candidate without watching Fox News, then watching Fox News will not change that. In other words, Assumption A rules out the possibility that the level of distrust that a voter has on Fox News is so high that she takes actions based on the opposite of the messages Fox News delivers.

Since assumption A is a key assumption in the paper, we first clarify how much we can hope for with and without Assumption A.

Lemma 1. *We generally have*

$$\max\left[0, \frac{\mathbb{P}\{Y_i(1) = 1\} - \mathbb{P}\{Y_i(0) = 1\}}{1 - \mathbb{P}\{Y_i(0) = 1\}}\right] \leq \theta_{pr} \leq \min\left[\frac{\mathbb{P}\{Y_i(1) = 1\}}{1 - \mathbb{P}\{Y_i(0) = 1\}}, 1\right], \quad (3)$$

where the bounds are sharp in that θ_{pr} can be anything between the bounds without changing the marginals $\mathbb{P}\{Y_i(1) = 1\}$ and $\mathbb{P}\{Y_i(0) = 1\}$. Further, Assumption A holds if and only if $\theta_{pr} = \theta_{avg}$, where

$$\theta_{avg} := \frac{\mathbb{P}\{Y_i(1) = 1\} - \mathbb{P}\{Y_i(0) = 1\}}{1 - \mathbb{P}\{Y_i(0) = 1\}}. \quad (4)$$

Equation (3) is a consequence of the Fréchet–Hoeffding inequality on the probability of a joint event. Since the two potential outcomes are never observed simultaneously, all we can ever hope to identify is their marginals and Equation (3) expresses the (sharp) bounds on θ_{pr} in terms of the marginal probabilities of the potential outcomes.

⁸Throughout the discussion, we assume that T_i is correctly measured if it is observed. See Calvi, Lewbel, and Tommasi (2022), Nguimkeu, Denteh, and Tchernis (2019), and Ura (2018) for the issues of mismeasured treatment. Their subject matter is distinct from ours.

⁹In online Appendix A, we present a simple economic model that motivates Assumption A.

Suppose that there are some voters who have such a high level of distrust on Fox News that rather not watching Fox News would help them to take favorable actions to a Republican candidate but that those voters are only minority and on average we still have a strict stochastic dominance relationship between $Y_i(1)$ and $Y_i(0)$ (i.e., $\mathbb{P}\{Y_i(1) = 1\} > \mathbb{P}\{Y_i(0) = 1\}$). Then, the lower bound on θ_{pr} will be ensured to be nontrivial.

Assumption **A** is stronger than the stochastic dominance, but it delivers a stronger result. In fact, Assumption **A** is necessary and sufficient to express θ_{pr} in terms of the marginal probabilities of the counterfactual outcomes. In this case, the conditional probability θ_{pr} is the average treatment effect (ATE) divided by $\mathbb{P}\{Y_i(0) = 0\}$. Throughout the rest of the paper, we present most of our results by using Assumption **A** because it is not only convenient but also being biased or directional seems to be the nature of persuasive effort. However, we emphasize that θ_{avg} , the rescaled version of the ATE, is always a valid lower bound on θ_{pr} as Equation (3) shows.

The next assumption is concerned about the treatment assignment T_i and the instrument Z_i .

Assumption B (No Defiers and an Exogenous Instrument). *The binary treatment T_i has a threshold structure, i.e.,*

$$T_i = \mathbb{1}\{V_i \leq e(Z_i)\}, \quad (5)$$

where V_i is an unobserved random variable that is uniformly distributed on $[0, 1]$. The binary instrument Z_i is independent of $(Y_i(t), V_i)$ for $t = 0, 1$. Finally, we have $0 \leq e(0) < e(1) \leq 1$.

Assumption **B** is standard for causal inference using instrumental variables. The intent-to-treat (ITT), Z_i , is randomly assigned; however, T_i can be endogenous via the dependence between V_i and $Y_i(t)$. The function $e(\cdot)$ is the propensity score or, more descriptively in our context, it can be referred to as the *exposure rate*.

If $T_i(z)$ denotes counterfactual treatment when a binary instrument takes value z , then agent i is called a complier when $T_i(1) = 1$ and $T_i(0) = 0$; never-takers ($T_i(z) = 0$ for all z), always-takers ($T_i(z) = 1$ for all z), and defiers ($T_i(1) = 0, T_i(0) = 1$) are similarly understood. Under Assumption **B**, i is a complier if and only if $e(0) < V_i \leq e(1)$. Similarly, i is an always-taker when $V_i \leq e(0)$, and she is a never-taker when $V_i > e(1)$. Therefore, under Assumption **B**, there are only three groups in the population: i.e., always-takers, never-takers and compliers. Indeed, as Vytlacil (2002) has shown, the threshold structure in Equation (5) is equivalent to assuming the absence of defiers, which is a popular assumption in econometrics to identify the local average treatment effect.

In the following two subsections, we present identification results for θ_{pr} , which is the same as θ_{avg} under assumption **A**. Section 3.2 covers the simplest case, where everybody complies so that there is no difference between the actual treatment and the intent to treat (ITT), i.e., $T_i = Z_i$ for each i : this case is referred to as the *sharp persuasion design*, where there is no distinction among different data scenarios. In

this case, not surprisingly, θ_{avg} is point identified from the distribution of Y_i given Z_i . However, when T_i and Z_i are different, which we call the *fuzzy persuasion design*, we have only partial identification of θ_{avg} , where each of the three data scenarios become relevant. Before we move on, we define

$$\theta_L := \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}{1 - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}, \quad (6)$$

provided that $\mathbb{P}(Y_i = 1 \mid Z_i = 0) < 1$: θ_L is an identified parameter from the distribution of Y_i given Z_i . It turns out that first, θ_L is equal to θ_{avg} under the sharp persuasion design, and second, it is the sharp lower bound on θ_{avg} in the fuzzy persuasion design regardless of which of the three data scenarios applies.

3.2. The Sharp Persuasion Design.

Theorem 1. *Suppose that Assumptions A and B hold. If $e(1) - e(0) = 1$ (i.e., $T_i = Z_i$ with probability one), then for $z \in \{0, 1\}$, we have $\mathbb{P}\{Y_i(z) = 1\} = \mathbb{P}(Y_i = 1 \mid Z_i = z)$, and hence $\theta_{\text{pr}} = \theta_{\text{avg}} = \theta_L$.*

The condition of $e(1) - e(0) = 1$ means that everybody is a complier, and hence there is essentially no difference between T_i and Z_i ; thus, the sharp design is equivalent to a situation where T_i is observed and randomized. However, this is rather an exceptional situation in social sciences. The key identification question should be how far we can go when the design is not sharp (i.e., not everybody is a complier). We answer this question in the following subsection.

Without Assumption A, the general sharp identified bounds on θ_{pr} are given by

$$\max\{0, \theta_L\} \leq \theta_{\text{pr}} \leq \min\left\{\frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1)}{1 - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}, 1\right\}. \quad (7)$$

Therefore, even in the sharp persuasion design, θ_{pr} is only partially identified and its sharp bounds can be trivial without Assumption A. Even so, it is worth noting that θ_L remains a valid lower bound.

3.3. The Fuzzy Persuasion Design. In the fuzzy design, the three scenarios of data availability we mentioned earlier become pertinent.

3.3.1. Identification with the Joint Distribution of (Y_i, T_i, Z_i) . Even the full joint distribution of (Y_i, T_i, Z_i) does not point-identify the ATE. Recall that those with $Z_i = 0$ and $T_i = 0$ comprise the compliers and the never-takers, while those with $Z_i = 0$ and $T_i = 1$ are the always-takers. Similarly, those with $Z_i = 1$ and $T_i = 0$ are the never-takers, while those with $Z_i = 1$ and $T_i = 1$ consist of the compliers and the always-takers. Therefore, these four cases correspond to different subpopulations, and the only subpopulation that we can study for both $T_i = 0$ and $T_i = 1$ in common is that of compliers, which explains why the Wald statistic estimates the LATE, not ATE. For the same reason, θ_{avg} cannot be point identified; however, we can derive its sharp bounds.

Assumption C (Full Observability). *The joint distribution of (Y_i, T_i, Z_i) is known, where $\mathbb{P}(Y_i = 0 \mid Z_i = 0) > 0$.*

Theorem 2. *Suppose that Assumptions A to C are satisfied. Then, the sharp identified interval of $\theta_{pr} = \theta_{avg}$ is given by $[\theta_L, \theta_U]$, where θ_L is given in Equation (6) and*

$$\theta_U := \frac{\mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) + 1 - e(1)}{1 - \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0)}.$$

To prove Theorem 2, we first derive the sharp identified bounds for $\mathbb{P}\{Y_i(1) = 1\}$ and $\mathbb{P}\{Y_i(0) = 1\}$ separately; we denote them by the intervals $[m_a, M_a]$ and $[m_b, M_b]$, respectively. These bounds are special cases of Manski and Pepper (2000) under the MTR assumption coupled with the exogeneity of the instrument. Then, letting $a := \mathbb{P}\{Y_i(1) = 1\}$ and $b := \mathbb{P}\{Y_i(0) = 1\}$, we obtain the upper bound of the identified interval of θ_{avg} by solving

$$\max_{a,b} \frac{a-b}{1-b} \quad \text{subject to} \quad a \in [m_a, M_a], b \in [m_b, M_b], a \geq b, \quad (8)$$

while the lower bound can be found by doing minimization instead of maximization. We then appeal to continuity and the intermediate value theorem for the sharpness result.

It is proved in online Appendix I that $m_a = \mathbb{P}(Y_i = 1 \mid Z_i = 1)$ and $M_b = \mathbb{P}(Y_i = 1 \mid Z_i = 0)$. Therefore, an examination of (8) reveals that the lower bound is attained when $a = m_a$ and $b = M_b$. To develop intuition behind Theorem 2, we discuss what $a = m_a$ and $b = M_b$ means, for which the behavior of the never-takers and that of the always-takers matter. Since $m_a = \mathbb{P}\{Y_i(1) = 1, T_i = 1 \mid Z_i = 1\} + \mathbb{P}\{Y_i(0) = 1, T_i = 0 \mid Z_i = 1\}$, we know that $a = m_a$ holds when $\mathbb{P}\{Y_i(0) = 1, T_i = 0 \mid Z_i = 1\} = \mathbb{P}\{Y_i(1) = 1, T_i = 0 \mid Z_i = 1\}$. Here, the event $Z_i = 1, T_i = 0$ means that i is a never-taker, because defiers are assumed to be non-existent. Therefore, $a = m_a$ means that the treatment has no effect on the group of never-takers, unless there are no never-takers at all. Similarly, $b = M_b$ holds when $\mathbb{P}\{Y_i(1) = 1, T_i = 1 \mid Z_i = 0\} = \mathbb{P}\{Y_i(0) = 1, T_i = 1 \mid Z_i = 0\}$. Since $Z_i = 0, T_i = 1$ means that i is an always-taker, we know that $b = M_b$ holds when the treatment does not affect the behavior of the always-takers, unless there are no always-takers at all. Therefore, θ_{avg} , which is the same the persuasion rate θ_{pr} under assumption A, is smallest when there are null treatment effects for both the never-takers and the always-takers.

Intuition for the upper bound can also be obtained by considering the non-complier groups. The upper bound corresponds to the case where $a = M_a$ and $b = m_b$, where it is shown in Appendix I that $M_a = \mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) + 1 - e(1)$ and $m_b = \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0)$. Here, note that $a = M_a$ is equivalent to $\mathbb{P}\{Y_i(1) = 0, T_i = 0 \mid Z_i = 1\} = 0$, and $b = m_b$ is to $\mathbb{P}\{Y_i(0) = 1, T_i = 1 \mid Z_i = 0\} = 0$. Therefore, we can see that the persuasion rate θ_{avg} equals the upper bound when every never-taker has $Y_i(1) = 1$ and none of the always-taker has $Y_i(0) = 1$: e.g., all those who never watch Fox News (whether it is available or not) would actually have voted for a Republican candidate if they had watched it and

all those who always watch Fox News would not have voted for a Republican without watching the channel.

The bounds in Theorem 2 shrink to a singleton as $(e(0), e(1))$ approaches $(0, 1)$, which is consistent with the result in Theorem 1. Also, it is worth noting that the lower bound θ_L only depends on the distribution of (Y_i, Z_i) : observing T_i along with (Y_i, Z_i) helps only for the upper bound. If $e(1)$ is too small, then the upper bound will not be very informative: θ_U converges to 1 as $e(1)$ approaches 0; that is, if nobody reads a newspaper when they receive free subscriptions, then we do not learn much about how “persuading” the newspaper is. However, even if $e(1)$ approaches 1, the upper bound does not necessarily shrink to the lower bound; e.g., we do not necessarily pin down the persuasion rate of reading the newspaper even if everybody who has free subscriptions actually reads it.

We now establish partial identification of θ_{pr} without Assumption A, i.e., $\theta_{pr} \neq \theta_{avg}$, for the sake of completeness. Let $NT = \{V_i > e(1)\}$ and $AT = \{V_i \leq e(0)\}$ be the event of i being a never-taker and an always-taker, respectively.

Theorem 3. *Suppose that Assumptions B and C are satisfied.*

(i) *If $\mathbb{P}(Y_i = 0 \mid Z_i = z) > 0$ for $z = 0, 1$, then the sharp identified interval of θ_{pr} is given by*

$$\max\left\{0, \frac{\mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) - e(0)}{1 - \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) - e(0)}\right\} \\ \leq \theta_{pr} \leq \min\left\{\frac{\mathbb{P}(Y = 1, T = 1 \mid Z = 1) + 1 - e(1)}{1 - \mathbb{P}(Y = 1, T = 0 \mid Z = 0)}, 1\right\}.$$

(ii) *If $\mathbb{P}(Y_i = 0 \mid Z_i = 0) > 0$ and $\mathbb{P}\{Y_i(1) = 1, \mathcal{E}\} \geq \mathbb{P}\{Y_i(0) = 1, \mathcal{E}\}$ for $\mathcal{E} \in \{NT, AT\}$, then the sharp identified interval of θ_{pr} is given by*

$$\max\{0, \theta_L\} \leq \theta_{pr} \leq \min\left\{\frac{\mathbb{P}(Y = 1, T = 1 \mid Z = 1) + 1 - e(1)}{1 - \mathbb{P}(Y = 1, T = 0 \mid Z = 0)}, 1\right\}.$$

Theorem 3 shows that θ_L continues to be the sharp lower bound, provided that $\theta_L \geq 0$, i.e., $\mathbb{P}(Y_i = 1 \mid Z_i = 1) \geq \mathbb{P}(Y_i = 1 \mid Z_i = 0)$, and a stochastic dominance condition holds for the never-takers and always-takers. However, the upper bound will be larger than that in Theorem 2 in general.

3.3.2. Identification with the Knowledge of the Exposure Rates. As in the case of DK, the researcher may not directly observe T_i along with (Y_i, Z_i) but may have auxiliary data from which the exposure rates $e(1)$ and $e(0)$ can be estimated.¹⁰ In this case, the sharp identified bounds on θ become generally wider than those of Theorem 2.

Assumption D (Observability of Two Marginals). *Only the distribution of (Y_i, Z_i) and the exposure rates $\{e(0), e(1)\}$ are known, where $\mathbb{P}(Y_i = 0 \mid Z_i = 0) > 0$.*

¹⁰The case in which the outcome and the treatment are separately observed belongs to an identification problem called the *ecological inference problem*. For instance, Cross and Manski (2002) and Manski (2018) discuss bounding a “long regression” by using information from a “short regression”. Their substantive concerns are distinct from ours.

Theorem 4. *Suppose that Assumptions A, B and D are satisfied. Then, the sharp identified interval of $\theta_{pr} = \theta_{avg}$ is given by $[\theta_L, \theta_{U_e}]$, where θ_L is given in Equation (6) and*

$$\theta_{U_e} := \frac{\min\{1, \mathbb{P}(Y_i = 1 | Z_i = 1) + 1 - e(1)\} - \max\{0, \mathbb{P}(Y_i = 1 | Z_i = 0) - e(0)\}}{1 - \max\{0, \mathbb{P}(Y_i = 1 | Z_i = 0) - e(0)\}}. \quad (9)$$

Therefore, the upper bound in this case is nontrivial if and only if $e(1) > \mathbb{P}(Y_i = 1 | Z_i = 1)$.¹¹ Note that it is the relative size of the take-up rate $e(1)$ (e.g., the probability of reading a newspaper when a free subscription to it is offered) that determines how much we can hope to learn about the persuasion rate. For example, if the probability of watching Fox News is too small relative to the probability of voting for a Republican candidate when Fox News was introduced in the local cable, then it becomes difficult to pin down how successfully Fox News persuaded their audience to vote for a Republican candidate. Also, it is worth noting that $e(0) = 0$ is not uncommon as Table 1 and Section 6 show. In this case, the maximum in the expression of the upper bound is unnecessary. Intuition for the lower bound is the same as the case of Theorem 2, because θ_L requires only the distribution of Y_i given Z_i .

3.3.3. *Identification with No Information Associated with T_i .* The final scenario is the least informative one, where T_i is not observed at all. This is an almost trivial case, but we state it in a separate theorem for the sake of completeness.

Assumption E (Limited Observability). *No information associated with T_i is available (i.e., the distribution of (Y_i, Z_i) is all that is known), where $\mathbb{P}(Y_i = 0 | Z_i = 0) > 0$.*

Theorem 5. *Suppose that Assumptions A, B and E are satisfied. Then, the sharp bound of $\theta_{pr} = \theta_{avg}$ is given by $[\theta_L, 1]$, where θ_L is given in Equation (6).*

The lower bound from Theorem 4 depends only on the distribution of (Y_i, Z_i) , and therefore θ_L continues to be the lower bound in this case as well. Further, since no information for $e(1)$ and $e(0)$ is available, it suffices to note that the upper bound in Theorem 4 equals one whenever $e(0) > \mathbb{P}(Y_i = 1 | Z_i = 0)$ and $e(1) < \mathbb{P}(Y_i = 1 | Z_i = 1)$.

4. THE LOCAL AND MARGINAL PERSUASION RATES

In this section, we consider rates of persuasion on other subpopulations that have been considered in econometrics: i.e.,

$$\begin{cases} \theta_{\text{local}} := \mathbb{P}\{Y_i(1) = 1 | Y_i(0) = 0, e(0) < V_i \leq e(1)\}, \\ \theta_{\text{marginal}}(v) := \mathbb{P}\{Y_i(1) = 1 | Y_i(0) = 0, V_i = v\} \text{ for } 0 < v < 1, \end{cases} \quad (10)$$

¹¹The trivial case can occur in applications. See Table 1 for such cases.

provided that the conditional probabilities are well-defined: θ_{local} is the persuasion rate for the compliers (e.g., [Imbens and Angrist, 1994](#)), whereas $\theta_{\text{marginal}}(v)$ is for the subpopulation such that $V_i = v$ (e.g., [Heckman and Vytlacil, 2005](#)).

First, we obtain identification results for θ_{local} under the three sampling scenarios in the fuzzy persuasion design. The first step for this purpose is to note that the same reasoning as Lemma 1 yields

$$\theta_{\text{local}} = \frac{\mathbb{P}\{Y_i(1) = 1 \mid e(0) < V_i \leq e(1)\} - \mathbb{P}\{Y_i(0) = 1 \mid e(0) < V_i \leq e(1)\}}{1 - \mathbb{P}\{Y_i(0) = 1 \mid e(0) < V_i \leq e(1)\}}, \quad (11)$$

where the numerator is the LATE, which has received great attention in the econometrics literature (see [Deaton, 2010](#); [Heckman, 2010](#); [Imbens, 2010](#), for a debate). The denominator that rescales the LATE is also conditioned on the same subpopulation of the compliers.

Theorem 6. *Suppose that Assumptions A and B are satisfied.*

(i) *Under Assumption C, θ_{local} is point identified by $\theta_{\text{local}} = \theta^*$, where*

$$\theta^* := \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}{\mathbb{P}(Y_i = 0, T_i = 0 \mid Z_i = 0) - \mathbb{P}(Y_i = 0, T_i = 0 \mid Z_i = 1)}.$$

(ii) *Under Assumption D, the sharp identified interval of θ_{local} is given by $[\theta_L^*, 1]$, where*

$$\theta_L^* := \max\left\{\theta_L, \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}{e(1) - e(0)}\right\}.$$

(iii) *Under Assumption E, the sharp identified interval of θ_{local} coincides with that of $\theta_{pr} = \theta_{\text{avg}}$, i.e., $[\theta_L, 1]$.*

Recall that the identification of the LATE requires the joint distribution of (T_i, Z_i) and that of (Y_i, Z_i) separately but not the full joint distribution of (Y_i, T_i, Z_i) . Unlike the LATE, the point identification in Theorem 6(i) demands the knowledge of the joint distribution of (Y_i, T_i, Z_i) .¹² Theorem 6(ii) shows that this requirement is not only sufficient but also necessary to achieve the point identification of θ_{local} .

Just like the LATE, it may be contentious whether or not θ_{local} should be the parameter of interest, because the compliers are concerned with an unidentified subgroup of the population. However, we take a practical view that the identification results on θ_{local} can complement the results obtained in Section 3.

The local persuasion rate θ_{local} represents the average persuasive effect for a population that is different from the entire population. Given this caveat, it is interesting to note that, in Theorem 6(ii), the upper bound on θ_{local} is always trivial in contrast to θ_{avg} , but the lower bound of θ_{local} can never be worse than that of θ_{avg} . Therefore, in principle, the length of the identified interval of θ_{avg} can be smaller than that of θ_{local} . If T_i is not observed at all, then there is no advantage in focusing on the compliers. Theorem 6(iii) confirms the intuition that the bounds for θ_{local} are identical to those for θ_{avg} if the distribution of (Y_i, Z_i)

¹²The denominator of Equation (11) requires that we know the marginal distribution of $Y_i(0)$ for the compliers. [Imbens and Rubin \(1997\)](#) show that the marginal distributions of $Y_i(1)$ and $Y_i(0)$ for the compliers are identified if the joint distribution of (Y_i, T_i, Z_i) is known; however, they did not consider the local persuasion rate.

is the only piece of information available. This corresponds to an uninteresting case for θ_{local} though, as we have no information on compliers.

Data requirements for the identification of $\theta_{\text{marginal}}(v)$ are generally quite demanding: e.g., a continuous instrument is needed. However, identifying $\theta_{\text{marginal}}(v)$ for various values of v can open up the possibility of point identification of θ_{avg} . Therefore, it is worth understanding what is sufficient for the identification of θ_{marginal} .

If Y_i and T_i are jointly observed along with a *continuous* instrument Z_i , then $\theta_{\text{marginal}}(v)$ can be point identified as in Heckman and Vytlacil (2005) and Carneiro, Heckman, and Vytlacil (2011). Examples of continuous instruments can be found in the literature on the media effects on voting. For instance, Enikolopov, Petrova, and Zhuravskaya (2011) and DellaVigna, Enikolopov, Mironova, Petrova, and Zhuravskaya (2014) use the signal strength of NTV and Serbian radio as instruments, respectively; in both of the papers, (Y_i, T_i, Z_i) are jointly observed. The following assumption describes the situation in which we can obtain point identification of $\theta_{\text{marginal}}(v)$. We use the standard results in the literature (e.g. Heckman and Vytlacil, 2005) for the subsequent theorem.

Assumption F (Marginal Treatment Effects). (i) *The joint distribution of (Y_i, T_i, Z_i) is known.*

(ii) *T_i has the threshold structure in Equation (5), where V_i is uniformly distributed on $[0, 1]$, and Z_i is independent of $(Y_i(t), V_i)$ for $t = 0, 1$.*

(iii) *The distribution of $e(Z_i)$ is absolutely continuous with respect to Lebesgue measure.*

Theorem 7. *Suppose that Assumptions A and F are satisfied. Then, for v such that v is in the interior of the support of $e(Z_i)$, $\theta_{\text{marginal}}(v)$ is point identified by*

$$\theta_{\text{marginal}}(v) = \frac{\partial \mathbb{P}\{Y_i = 1 \mid e(Z_i) = e\} / \partial e|_{e=v}}{1 + \partial \mathbb{P}\{Y_i = 1, T_i = 0 \mid e(Z_i) = e\} / \partial e|_{e=v}}, \quad (12)$$

provided that $\mathbb{P}\{Y_i = 1 \mid e(Z_i) = e\}$ and $\mathbb{P}\{Y_i = 1, T_i = 0 \mid e(Z_i) = e\}$ are continuously differentiable with respect to e .

Similarly to the case of θ_{avg} or θ_{local} , Assumption A enable us to rewrite $\theta_{\text{marginal}}(v)$ as $\mathbb{E}\{Y_i(1) - Y_i(0) \mid V_i = v\} / \mathbb{P}\{Y_i(0) = 0 \mid V_i = v\}$; i.e., $\theta_{\text{marginal}}(v)$ is a rescaled version of the marginal treatment effect of Heckman and Vytlacil (2005). Theorem 7 is a direct consequence of that.

Theorem 7 does not consider the other two scenarios of data availability. This is mainly because continuous instruments are relatively infrequent in the context of persuasion, and we are not aware of any applications where continuous instruments are available while the outcome and treatment are not jointly observed.

If the support of the exposure rate $e(Z_i)$ is equal to the unit interval $[0, 1]$, then Theorem 7 shows the identification of $\theta_{\text{marginal}}(v)$ for all v in the unit interval. Then, we can use $\theta_{\text{marginal}}(v)$ to construct different policy-oriented quantities as in Heckman and Vytlacil (2005) and Carneiro, Heckman,

and Vytlačil (2011). For instance, the persuasion rate of the entire population can be obtained by $\int_0^1 \theta_{\text{marginal}}(v) dF\{v \mid Y_i(0) = 0\}$, which is equal to

$$\theta_{\text{avg}} = \frac{\int_0^1 \partial \mathbb{P}(Y_i = 1 \mid e(Z_i) = e) / \partial e|_{e=v} dv}{1 + \int_0^1 \partial \mathbb{P}\{Y_i = 1, T_i = 0 \mid e(Z_i) = e\} / \partial e|_{e=v} dv}$$

by Bayes' theorem.

5. DISCUSSION

In this section we articulate the relationships between θ_{avg} , θ_{local} , and DK's measures f and \tilde{f} defined in Section 2, and we summarize the main takeaways of our identification results. We assume that Assumptions A and B hold throughout this section, so we have $\theta_{\text{pr}} = \theta_{\text{avg}}$. Also, in order to be consistent with the identification analysis, we work with the population versions of f and \tilde{f} : i.e.,

$$f \xrightarrow{p} \theta_{DK} := \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}{e(1) - e(0)} \frac{1}{1 - \mathbb{P}\{Y_i(0) = 1\}}, \quad (13)$$

$$\tilde{f} \xrightarrow{p} \tilde{\theta}_{DK} := \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}{e(1) - e(0)} \frac{1}{1 - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}. \quad (14)$$

First, neither θ_{DK} nor $\tilde{\theta}_{DK}$ is generally equal to the persuasion rate θ_{avg} , or that for the compliers, θ_{local} , which is clear from the fact that

$$\mathbb{P}\{Y_i(0) = 0\} \theta_{DK} = \mathbb{P}\{Y_i = 0 \mid Z_i = 0\} \tilde{\theta}_{DK} = \mathbb{P}\{Y_i(0) = 0 \mid e(0) < V_i \leq e(1)\} \theta_{\text{local}}$$

is equal to the LATE, while $\mathbb{P}\{Y_i(0) = 0\} \theta_{\text{avg}}$ is equal to the ATE.¹³ For example, θ_{DK} rescales the LATE with an unconditional probability and hence it does not render a well-defined conditional probability in general. Similarly, $\tilde{\theta}_{DK}$ is not necessarily a 'rate' in spite of the rescaling factor. So, making comparisons across different studies based on θ_{DK} or $\tilde{\theta}_{DK}$ can be misleading, although it is a common practice (e.g., DellaVigna and Gentzkow, 2010).

There are some special cases of exception though: (i) everybody is a complier as in the sharp persuasion design, (ii) T_i is independent of the potential outcomes $Y_i(t)$ for $t = 0, 1$ conditional on Z_i , or (iii) there is no heterogeneity in the treatment effect in that $Y_i(1) - Y_i(0)$ is a constant. That is, in cases (i) and (ii), there is no endogeneity issue, whereas in case (iii), no one is affected by the persuasive message (i.e., $Y_i(1) - Y_i(0) = 0$ for all i), or everybody is persuaded (i.e. $Y_i(1) - Y_i(0) = 1$ for all i). Note that the LATE is the same as the ATE when any of the three cases applies. As a result, we have that in case (i), $\theta_{DK} = \tilde{\theta}_{DK} = \theta_{\text{avg}} = \theta_{\text{local}}$ holds; in case (ii), $\theta_{DK} = \theta_{\text{avg}} = \theta_{\text{local}} \leq \tilde{\theta}_{DK}$; and, in case (iii), $\theta_{DK} = \theta_{\text{avg}} = \theta_{\text{local}} = 1 \leq \tilde{\theta}_{DK}$ if everybody is persuaded, or $\theta_{DK} = \tilde{\theta}_{DK} = \theta_{\text{avg}} = \theta_{\text{local}} = 0$, unless any of them are ill-defined, if no one is affected by the informational treatment. Therefore, the feasible version \tilde{f} of DK's measure of persuasion does estimate the persuasion rate θ_{avg} in case (i), although,

¹³Recall that the population version of the Wald statistic is equal to the LATE under Assumption B.

as DK correctly pointed out, $\tilde{\theta}_{DK}$ does approximate θ_{DK} in case (ii) as well if either $e(0)$ or θ is close to zero.¹⁴

Our identification results in the previous sections show that θ_L is the sharp lower bound of the identified interval of θ_{avg} in the fuzzy design regardless of whether the full joint distribution of the outcome, treatment, and instrument is available or not. The parameter θ_L has been reported in the literature without understanding that it is the sharp lower bound on θ_{avg} . For instance, DellaVigna and Gentzkow (2010) extensively estimate $\tilde{\theta}_{DK}$ by using many examples but they report θ_L as a lower bound of $\tilde{\theta}_{DK}$ when T_i 's are unobserved and hence $e(1)$ and $e(0)$ are unknown. Our results show that θ_L is *always* a meaningful parameter, but $\tilde{\theta}_{DK}$ may not. Therefore, even when information about $e(0)$ and $e(1)$ is available, θ_L is a better parameter to estimate than $\tilde{\theta}_{DK}$.

Indeed, if the full joint distribution of (Y_i, T_i, Z_i) is available, then we recommend reporting $[\theta_L, \theta_U]$ along with θ^* ; these can be consistently estimated by their sample analogs. If (Y_i, Z_i) is observed with some auxiliary information for $e(0)$ and $e(1)$ available, then $[\theta_L, \theta_{U_e}]$ and $[\theta_L^*, 1]$ should be reported. If T_i is not observed at all, then the interval $[\theta_L, 1]$ is the best we can hope for to study either θ_{avg} or θ_{local} .

Note that θ_L should be estimated all the time; it only requires data on (Y_i, Z_i) . Because the actual T_i can be difficult to observe, researchers have used an extra micro-level survey to obtain auxiliary data on T_i , which seems quite costly. However, the value of an attempt to observe T_i can be limited, depending on which parameter the researcher wants to learn about. For instance, if the researcher cares about the persuasion rate of the entire population, then observing T_i does not add any information for the lower bound, while it can potentially improve the upper bound. If the group of compliers is of interest, then whether we observe T_i or not, and how we observe it, can be relevant issues; we have $\theta_L^* \geq \theta_L$ in the second data scenario and θ^* is point identified if (Y_i, T_i, Z_i) is jointly observed. If Z_i is continuously distributed, the value of observing (Y_i, T_i, Z_i) jointly increases dramatically as well. In summary, our identification analysis shows that the value of observing T_i depends crucially on which population the researcher is interested in.

In order to illustrate the difference between DK's measure and our bounds, we have calculated them in Table 1. We focus on the results reported in DellaVigna and Gentzkow (2010, their Table 1) when the outcome variable is voter turnout. We have chosen this type of study as the turnout is among the most studied outcome variables in the literature and it is naturally a binary measure. Table 1 provides estimates of $\mathbb{P}(Y_i = 1 \mid Z_i = z)$ and $e(z)$ for $z = 0, 1$, thereby enabling us to obtain the bounds based on theorem 4 and theorem 6 (ii). It can be seen that using DK's persuasion rates alone may lead to misleading conclusions because the bounds on θ_{avg} as well as those on θ_{local} are in fact wide. Moreover, the results in Table 1 suggest that identification power under Assumptions A, B and D in this example

¹⁴In case (ii), we have $\mathbb{P}(Y_i = 1 \mid Z_i = 0) = \mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 0) + \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) = \mathbb{P}\{Y_i(0) = 1\} + [\mathbb{P}\{Y_i(1) = 1\} - \mathbb{P}\{Y_i(0) = 1\}]e(0)$.

is limited, especially for the upper bounds on θ_{avg} and θ_{local} . We further illustrate these points with empirical examples in Section 6.

Finally, since the parameters are partially identified, inference should also account for that. The method proposed by [Stoye \(2009\)](#) is useful for that purpose, at least in the most favorable data scenario, in which case the sample analog principle and the delta method show that we can construct the estimators $\hat{\theta}_L$ and $\hat{\theta}_U$ that are asymptotically jointly normal. Therefore, by [Stoye \(2009\)](#), a $(1 - \alpha)$ confidence interval for θ_{avg} can be obtained by $[\hat{\theta}_L - c_\alpha \hat{\sigma}_L, \hat{\theta}_U + c_\alpha \hat{\sigma}_U]$, where $\hat{\sigma}_L$ and $\hat{\sigma}_U$ are the estimated standard errors of $\hat{\theta}_L$ and $\hat{\theta}_U$, respectively, and c_α is chosen by solving

$$\Phi\left(c_\alpha + \frac{\hat{\Delta}}{\max(\hat{\sigma}_L, \hat{\sigma}_U)}\right) - \Phi(-c_\alpha) = 1 - \alpha,$$

where Φ is the distribution function of the standard normal and $\hat{\Delta}$ is the estimated length of the identified interval.

The second data scenario is slightly more complicated, because θ_{U_e} and θ_L^* contain the min or max function that is not smooth; so, the delta method does not apply. In online Appendices [F](#) and [G](#), we propose a two-step method for inference to overcome this problem, which we have applied to the empirical example we discuss in Section 6. In the third data scheme, confidence intervals for θ_{avg} and θ_{local} always coincide, and they can be obtained by using a one-side critical value on $\hat{\theta}_L$. Specifically, they are given by $[\hat{\theta}_L - z_{1-\alpha} \hat{\sigma}_L, 1]$, where $z_{1-\alpha}$ is the $(1 - \alpha)$ quantile of the standard normal distribution. Online Appendices [F](#) and [G](#) provide a more detailed discussion on inference. Furthermore, see online Appendix [H](#) for semiparametrically efficient estimation of the two key parameters, i.e., θ_L and θ^* , when exogenous covariates X_i are present and integrated out.

In sum, this paper clarifies identification issues when we insert exposure as a choice variable and employ a proper causal framework that is used in policy evaluation to model two causal links (i.e., $Z_i \rightarrow T_i$ and $T_i \rightarrow Y_i$).¹⁵

6. EMPIRICAL EXAMPLES

6.1. Effects of Uncensored Media. In this subsection, we revisit [Chen and Yang \(2019, CY hereafter\)](#), who conducted a field experiment in China to measure the effects of providing students with internet access to the uncensored media on various outcome variables. Excluding the existing users, the subjects in their experiments consist of four groups: (i) the control group; (ii) the control group who were encouraged to visit foreign news websites blocked by the Great Firewall; (iii) students who received free access to uncensored internet; and (iv) students who received both the access and encouragement treatments. They followed the subjects over 18 months to collect outcomes on media-related behaviors, beliefs, and attitudes among other things. It turns out that there were no differences between groups (i) and (ii)

¹⁵We are grateful to an anonymous referee who provided us with insightful comments.

and the effects were the largest for group (iv), i.e., the access plus encouragement group (the *Group-AE* students from now on). To benchmark their findings, CY computed DK’s measure of persuasion for the *Group-AE* students (see online Appendix Table A.13 of their paper for the details) and commented that “their estimated persuasion rates are of a similar magnitude to those found in authoritarian regimes that typically have highly regulated media markets” (see pp. 2323–24 in CY).

In this section, we use the data from CY to illustrate how the common practice of reporting DK’s measure can lead to misleading conclusions by contrasting DK’s measure of persuasion with our proposed approaches. As in CY, we focus on the *Group-AE* students. That is, $Z_i = 1$ if the i th subject is randomly assigned to the *Group-AE* group, and $Z_i = 0$ if the i th subject is randomly assigned to the control or control-encouragement group, while dropping the access only group and the existing users. In CY’s two-stage least squares analysis (Table 3 in their paper), the treatment variable is $T_i = 1$ if the i th subject is an active user of the censorship circumvention tool and $T_i = 0$ otherwise. We use the same treatment variable in our analysis. To replicate the results in CY in a representative but succinct way, we focus on the 11 outcome variables listed in Panel A of online Appendix Table A.13 in their paper. They represent media-related behaviors, beliefs, and attitudes and are transformed to binary variables by CY.

Recall the population version of DK’s measure f ; see Equation (13). In their online Appendix Table A.13, CY measure $\mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0)$ by the *intent-to-treat* effects of the *Group-AE* assignment, and approximate $\mathbb{P}\{Y_i(0) = 1\}$ using variables collected at the time of the baseline survey or by the estimates of $\mathbb{P}(Y_i = 1 \mid Z_i = 0)$ as in $\tilde{\theta}_{DK}$, if the former is unavailable. Table 2 in CY and the data provided by CY indicate that the change in the exposure rate is 45.5% if the treatment status is measured by being active users, and therefore we use $e(1) - e(0) = 0.455$ for our subsequent calculations.

Table 2 summarizes the empirical results. In Column (2) of Table 2, we recompute CY’s persuasion rates for the 11 outcome variables listed in Panel A of online Appendix Table A.13 in their paper. As we explained above, these estimates are based on $e(1) - e(0) = 0.455$. Out of the 11 outcome variables, the median persuasion rate is 101%, and therefore these “persuasion rates” cannot be understood as conditional probabilities. Columns (3) and (4) report our estimates of the average and local persuasion rates along with the 95% confidence intervals (in curly braces) that were obtained via 10,000 bootstrap replications. Our estimates show that (i) the average persuasion rates are only partially identified and the widths of the identified intervals are substantial for most of the outcomes, (ii) the point-identified local persuasion rates are contained by the identified intervals for the average persuasion rates and are typically closer to the upper end points of the intervals. The persuasive effects are of a relatively large magnitude in that the smallest lower end point of the confidence interval for the average persuasion is 16%. However, the original estimates overstate the magnitude by a substantial factor and mask under-identification of average persuasion rates. In short, we find that the subjects in the experiments

responded to exposure to uncensored internet highly heterogeneously, indicating that it is important to go beyond the benchmark measures of DK type.

6.2. Effects of Political News. We now illustrate our proposed methods by using data from Gerber, Karlan, and Bergan (2009, GKB hereafter), who report findings from a field experiment to measure the effect of political news. We have chosen this example because it contains a credible binary instrument from the field experiment and we can also illustrate all of the three sampling scenarios as well as the case of nonbinary outcomes; for the theory on the multinomial outcome case, see online Appendix B. In GKB, there are three statuses in the intention to treat: a control group, an offer of free subscription to *The Washington Post*, and one to *The Washington Times*. To illustrate the usefulness of our paper, we focus on *The Washington Post* and drop all observations from *The Washington Times* subscription. That is, $Z_i = 1$ if the i th individual received free subscription to *The Washington Post*, and $Z_i = 0$ if not.

Focusing on the ITT analysis, GKB have reported ITT estimates for various outcomes Y_i . DellaVigna and Gentzkow (2010) compute persuasion rates for GKB, for which they simply set $T_i = 1$ if the i th individual opted into the free subscription and $T_i = 0$ if they opted out of it.¹⁶ In this section, for the purpose of illustrating our identification results, we consider a different treatment variable: $T_i = 1$ if the i th individual read a newspaper at least several times per week and $T_i = 0$ otherwise, which is a variable that GKB kept track of in a follow-up survey. Therefore, the relevant treatment we consider differs from that of DellaVigna and Gentzkow (2010), but it is whether individuals have *actually* read the newspaper or not. The outcome variables we consider are as follows. For the binary case, $Y_i = 1$ if the i th individual reported voting for the Democratic candidate in the 2005 gubernatorial election, and $Y_i = 0$ if the subject did not vote for the Democratic candidate or did not vote at all. For the multinomial case, not voting at all is treated as an outside option. We use only a subsample of the GKB data with those who responded to the follow-up survey to use information on (Y_i, T_i, Z_i) jointly. After dropping observations for *The Washington Times* subscription and removing missing data, we summarize the GKB data in Table 3. Although the joint distribution of (Y_i, T_i, Z_i) is observed in this example, we also consider using the two marginals of (Y_i, Z_i) and (T_i, Z_i) separately, to make a comparison. The estimates are summarized in Table 4. Because the size of the sample extract we use is relatively modest ($n = 701$) for an interval-identified object, we report the 80% confidence intervals obtained by the inference methods described in Section 5 as well as in online Appendices F and G.

First, we discuss the case where the full joint distribution of (Y_i, T_i, Z_i) is used. In this data scenario, the average effect of persuasion by reading the newspaper is bounded between 7% and 63%. In contrast, the persuasion rate for the group of compliers is point-estimated by 81%. It is interesting to note that the estimate of θ_{local} is so large that it is greater than the upper bound of θ_{avg} . This suggests that individuals

¹⁶We provide the empirical results of bound analysis using the opting-into-the-free-subscription treatment variable in online Appendix C.

are highly heterogeneous in this example, indicating that $\tilde{\theta}_{DK}$ might not be a well-defined conditional probability here. Indeed, the estimate of $\tilde{\theta}_{DK}$ in Equation (14) is $\hat{\theta}_{DK} = 1.1027$, which is greater than one.

When the marginals of (Y_i, Z_i) and (T_i, Z_i) are used separately, the upper bound on θ_{avg} increases from 63% to 78%. Further, θ_{local} is no longer point estimated but we only know that it is bounded between 78% and 100%. This difference illustrates the loss of identification power if we do not observe the joint distribution of (Y_i, T_i, Z_i) .

Finally, we estimate the lower bound on the average persuasion rate by additionally conditioning on those who would vote even without reading the newspaper (see online Appendix B for details). The resulting lower bound on the average persuasion increases from 0.0707 (0.0289) to 0.0975 (0.0554), where the numbers in the parentheses are the left-end points of the 80% confidence intervals. Therefore, the (point identified) ITT effect is 5%, while the lower bound of the average persuasion rate is about 7%, or 10% if we further focus on those who would vote without reading the newspaper.

7. CONCLUSIONS

We have set up a simple econometric model of persuasion, introduced several parameters of interest, and analyzed their identification. The empirical examples in Sections 5 and 6 as well as the examples in online Appendices D and E demonstrate that the persuasive effects are highly heterogeneous in the settings of media and fundraising.

We have focused on the case of binary outcomes and binary treatments. In online Appendix B, we extend our analysis to nonbinary outcomes. If the outcome is nonbinary, then we can condition on those who would not choose the outside option without the treatment. For instance, suppose that we have three options of voting for a Republican, voting for a Democrat, or not voting at all. Then, the persuasive effect of a message supporting a Republican can be measured in a couple of different ways: focusing on those who would not have voted for a Republican without the message is one way and conditioning on those who would have voted for a Democrat (i.e., voted but not voted for a Republican) is the other. In the latter case, we show that the resulting lower bound is always no smaller than that of the binary outcome case.

In general, treatments are multivalued: unordered treatments (e.g., watching Fox News, CNN or MSNBC) and ordered treatments (e.g., numbers of hours watching Fox News) arise naturally in applications. It would be fruitful to build on recent developments in multivalued treatments (e.g., Heckman, Urzua, and Vytlačil, 2006; Heckman and Vytlačil, 2007b; Heckman, Urzua, and Vytlačil, 2008; Heckman and Pinto, 2018; Lee and Salanié, 2018) to investigate identification of persuasive effects. It would also be interesting to estimate deep parameters in an economic model of persuasion by using a more structural approach in the set-up of multivalued treatments. These are topics for future research.

TABLE 1. Persuasion Rates: Papers on Voter Turnout

Paper	$\hat{y}(1)$	$\hat{y}(0)$	$\hat{\epsilon}(1)$	$\hat{\epsilon}(0)$	\bar{f}	$[\hat{\theta}_L, \hat{\theta}_{U_e}]$	$[\hat{\theta}_L^*, 1]$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Green and Gerber (2000)	.472	.448	.279	0	.156	[.043,1]	[.086,1]
Green, Gerber, and Nickerson (2003)	.310	.286	.293	0	.115	[.034,1]	[.082,1]
Green and Gerber (2001)	.711	.660	.737	0	.204	[.150,.924]	[.150,1]
Green and Gerber (2001)	.416	.405	.414	0	.045	[.018,1]	[.027,1]
Gentzkow (2006)	.455	.435	.800	0	.044	[.035,.389]	[.035,1]
Gentzkow, Shapiro, and Sinkinson (2011)	.700	.690	.250	0	.129	[.032,1]	[.040,1]

^a The outcome variable is voter turnout except for [Gentzkow \(2006\)](#), where exposure to television discouraged voters to go to the polls. Thus, to have positive persuasive effects in all rows, the outcome variable for [Gentzkow \(2006\)](#) is not to vote.

^b In columns (1) and (2), $\hat{y}(z)$ denotes estimates of $\mathbb{P}(Y_i = 1 \mid Z_i = z)$ for $z = 1, 0$.

^c \bar{f} is the persuasion rate reported in [DellaVigna and Gentzkow \(2010\)](#), table 1).

^d $[\theta_L, \theta_{U_e}]$ and $[\theta_L^*, 1]$ are the sharp lower and upper bounds on the average and local persuasion rates, respectively, under Assumptions [A](#), [B](#) and [D](#).

^e The third row corresponds to the row in Table 1 of [DellaVigna and Gentzkow \(2010\)](#) under the treatment labeled “phone calls by youth vote”.

^f The fourth row corresponds to the row in Table 1 of [DellaVigna and Gentzkow \(2010\)](#) under the treatment labeled “phone calls 18–30-year-olds”.

^g Table 1 of [DellaVigna and Gentzkow \(2010\)](#) gives $\hat{\epsilon}(1) - \hat{\epsilon}(0)$ but $\hat{\epsilon}(0) \equiv 0$ in each row by study design.

TABLE 2. Persuasion Rates of Exposure to Uncensored Internet

(1)	(2)	(3)	(4)
Outcome	CY19	Persuasion Rates Average	Local
A.1.2 Ranked high: foreign websites	35.5	[21.9, 68.4] {15.6, 72.9}	41.0 {29.2, 51.1}
A.1.6 Freq. of visiting foreign websites for info.	168.5	[55.5, 76.6] {50.6, 80.2}	70.4 {64.6, 75.7}
A.2.1 Purchase discounted tool we offered	45.6	[22.7, 64.2] {19.3, 67.9}	38.7 {32.6, 44.9}
A.2.2 Purchase any tool	101.2	[49.5, 85.0] {45.5, 87.7}	76.8 {71.9, 81.6}
A.3 Valuation of access to foreign media outlets	121.7	[49.3, 82.1] {42.8, 85.8}	73.4 {65.8, 80.0}
A.4 Trust in non-domestic media outlets	158.1	[62.9, 85.8] {57.8, 89.0}	81.6 {76.2, 86.4}
A.5.1 Degree of censorship on domestic news outlets	82.3	[35.1, 73.5] {29.6, 77.4}	57.0 {48.7, 64.3}
A.5.2 Degree of censorship on foreign news outlets	114.6	[39.1, 74.4] {33.8, 78.2}	60.4 {53.1, 67.3}
A.6 Censorship unjustified	77.6	[31.6, 66.6] {23.5, 72.0}	48.6 {35.8, 59.2}
A.7.1 Domestic cens. driven by govt. policies	192.7	[75.0, 88.6] {64.4, 94.2}	86.8 {76.5, 94.4}
A.7.2 Foreign cens. driven by govt. policies	54.1	[36.9, 76.3] {19.2, 84.1}	61.0 {32.4, 77.4}

^a In Column (2), we recompute CY's persuasion rates for the 11 outcome variables listed in Panel A of online Appendix Table A.13 in their paper. For these estimates, we use $e(1) - e(0) = 0.455$.

^b Column (3) reports the bound estimates of the average persuasion rates in brackets.

^c Column (4) reports the point estimates of the local persuasion rates.

^d The 95% confidence intervals are given in curly braces, obtained via 10,000 bootstrap replications.

^e The persuasion rates are expressed in percentage.

TABLE 3. Summary Statistics of the GKB Data

<i>The Washington Post</i> ($Z_i = 1$)			
	Read a newspaper		Total
Voted for Democrat	$T_i = 0$	$T_i = 1$	
$Y_i = 0$	94	93	187
$Y_i = 1$	31	68	99
Total	125	161	286
Control ($Z_i = 0$)			
	Read a newspaper		Total
Voted for Democrat	$T_i = 0$	$T_i = 1$	
$Y_i = 0$	162	130	292
$Y_i = 1$	46	77	123
Total	208	207	415

The GKB data are used after dropping observations for *The Washington Times* subscription and removing missing data.

TABLE 4. Estimates of the Key Parameters

	(Y_i, T_i, Z_i)	(Y_i, Z_i) and (T_i, Z_i)	(Y_i, Z_i) only
ITT	0.0498 {0.0036, 0.0959}		
θ_{avg}	[0.0707, 0.6343] {0.0289, 0.6610}	[0.0707, 0.7832] {0.0286, 0.8143}	[0.0707, 1] {0.0288, 1}
LATE	0.7759 {0, 1}		[0.0498, 1] {0.0195, 1}
θ_{local}	0.8067 {0.1243, 1}	[0.7759, 1] {0.0069, 1}	[0.0707, 1] {0.0288, 1}

^a The first and third rows show the estimates of ITT and LATE, respectively. The second row corresponds to $[\hat{\theta}_L, \hat{\theta}_U]$, $[\hat{\theta}_L, \hat{\theta}_{U^*}]$, and $[\hat{\theta}_L, 1]$, respectively. The fourth row shows $\hat{\theta}^*$, $[\hat{\theta}_L^*, 1]$, and $[\hat{\theta}_L, 1]$, respectively.

^b 80% confidence intervals are given in curly braces.

^c The sharp identified interval for LATE is $[\text{ITT}, 1]$ when only the joint distribution of (Y_i, Z_i) is available. This is because $e(1) - e(0)$ is unknown and can be everywhere between 0 and 1 in this case.

 ONLINE APPENDICES TO “IDENTIFYING THE EFFECT OF PERSUASION”

INTRODUCTION

The online appendices of the paper are grouped into four parts. Part I includes additional results that are omitted from the main text. In Appendix A, we present a simple economic model to motivate our setup. Focusing on a binary treatment and a binary outcome, we formulate a model of persuasion within the framework of expected utility maximization. This formulation naturally leads to a potential outcome setup with a certain monotonicity restriction. In Appendix B, we give identification results with nonbinary outcomes.

Part II provides additional empirical examples. In Appendix C, we provide further discussion on Section 6 by analyzing the case where treatment is whether individuals opted into the free subscription to *The Washington Post*. In Appendix D, we revisit the empirical literature on the effects of news media on voting, where we apply our identification results to two published articles. In particular, when we revisit DellaVigna and Kaplan (2007, DK hereafter) using their original data, we find that the identification region for the persuasion rate θ_{avg} is between 1% and 100% and that the lower bound for the local persuasion rate θ_{local} is either 12% or 38%, depending on the specification of the fixed effects. These results suggest that the persuasive effect of Fox News is fairly large for compliers, i.e., those who would watch Fox News if and only if it is available via the local cable, but that DK’s data are uninformative about the general population. In Appendix E, we look at the literature on door-to-door fundraising and we illustrate the usefulness of our results by applying them to two published papers.

Part III deals with estimation and inference problems. In Appendix F, we explain methods for Inference on the average persuasion rate; in Appendix G, we describe methods for inference on the local persuasion rate. In Appendix H, we consider semiparametrically efficient estimation of the two key parameters, i.e., the lower bounds on θ_{avg} and θ_{local} , and we provide an empirical illustration.

Part IV, which is Appendix I, contains all the proofs.

PART I. ADDITIONAL THEORETICAL RESULTS

APPENDIX A. A MICROECONOMIC FOUNDATION FOR ASSUMPTION A

In this section, we consider a binary choice problem under binary states. The states are unknown to the agents at the time of the decision and they rely on their subjective belief about the states to make a decision. Suppose that two possible states are denoted by $S \in \mathcal{S} = \{\text{High}, \text{Low}\}$. Let $T_i \in \{0, 1\}$ indicate individual i ’s status of the informational treatment. Further, let $q_i(t)$ describe individual i ’s subjective belief about the state when T_i is set to $t \in \{0, 1\}$: i.e., $q_i(t) = \mathbb{P}(S = \text{High} \mid T_i = t, I_i)$, where I_i denotes all other information available to individual i . Table A1 describes the utility individual

i receives from each choice conditional on the state. The payoffs matrix in Table A1 is from Bergemann and Morris (2019, see matrix (5)).

TABLE A1. Utility by choice and state

	$S = \text{Low}$	$S = \text{High}$
Vote (1)	-1	$U_i \geq 0$
Not vote (0)	0	0

The utility from option 0 is normalized to be 0 for each state. Since the expected utility is all that matters for the decision, the utility from option 1 when the state is “low” is normalized to be -1 : the sign restrictions are to make the choice nontrivial. The utility term U_i is not observed by the econometrician.

Suppose that individual i maximizes expected utility. Then, individual i chooses option 1 if and only if expected utility, $-\{1 - q_i(t)\} + q_i(t)U_i$ for $t \in \{0, 1\}$, is positive with subjective belief $q_i(t)$ about the state. Therefore, when the informational treatment is set to be $t \in \{0, 1\}$, the potential outcome $Y_i(t)$ can be written as follows:

$$Y_i(t) = \mathbb{1}[-\{1 - q_i(t)\} + q_i(t)U_i \geq 0],$$

where $\mathbb{1}[\cdot]$ is the usual indicator function. We now make the following assumptions.

Assumption A1. U_i has a conditional density $f\{\cdot \mid q_i(0), q_i(1)\}$ such that $f\{u \mid q_i(0), q_i(1)\} > 0$ for all $u \in [0, \infty)$ with probability one.

Assumption A2. $q_i(0) \leq q_i(1)$ with probability one.

Assumption A1 says that U_i is continuously distributed given $q_i(0)$ and $q_i(1)$. However, it does not rule out the possibility that U_i and $q_i(t)$ are dependent on each other. Assumption A2 simply means that the informational treatment may shift an agent’s belief *only in one direction*.

Lemma A1. Under Assumption A1, Assumption A2 is equivalent to $Y_i(0) \leq Y_i(1)$ with probability one.

Therefore, Lemma A1 provides a microeconomic foundation for Assumption A.

APPENDIX B. AN EXTENSION TO NONBINARY OUTCOMES

B.1. The Setup and the sharp lower bound. When outcomes are not binary, one might want to treat an outside option differently. In the voting example, those who would go out and vote even without any informational treatment may be the relevant subpopulation to consider in defining the rate of persuasion. Below we formalize this idea.

Suppose that $\mathcal{S} = \{0, 1, -1\}$, where 0 is an outside option, 1 is the target action of persuasion, and -1 represents taking any other action. For instance, taking action 0 can mean that the agent does not vote at

all, whereas taking action 1 means that the agent votes for a candidate from party 1 and taking -1 means that the agent votes for a candidate from any other party. We then denote agent i 's potential outcomes by the vector of binary variables $Y_i(t) = (Y_{i0}(t), Y_{i1}(t), Y_{i,-1}(t))$ for $t \in \{0, 1\}$. Finally, we assume that the choices in \mathcal{S} are exclusive and exhaustive so that $\sum_{j \in \mathcal{S}} Y_{ij}(t) = 1$ for $t \in \{0, 1\}$.¹⁷ Similarly to the binary case, we impose monotonicity on the target action of persuasion (i.e., $Y_{i1}(1) \geq Y_{i1}(0)$ with probability one). The following assumption summarizes our set-up.

Assumption B1 (Multinomial Outcomes). $Y_{i1}(1) \geq Y_{i1}(0)$ and $\sum_{j \in \mathcal{S}} Y_{ij}(t) = 1$ for $t = 0, 1$ with probability one.

If we only focus on the target action $Y_{i1}(t)$, then the persuasion rate θ_{pr} can be defined as in Section 2 (i.e., $\theta_{\text{pr}} = \mathbb{P}\{Y_{i1}(1) = 1 \mid Y_{i1}(0) = 0\}$). However, if one wants to be explicit about the presence of the outside option, then conditioning on those who would not choose the outside option without the treatment seems appropriate to define the rate of persuasion; that is,

$$\theta_{\text{mult}} := \mathbb{P}\{Y_i(1) = (0, 1, 0) \mid Y_i(0) = (0, 0, 1)\} = \mathbb{P}\{Y_{i1}(1) = 1 \mid Y_{i0}(0) = 0, Y_{i1}(0) = 0\}, \quad (\text{B1})$$

provided that the conditional probability is well-defined: the second equality uses Assumption B1. Note that θ_{mult} is a different parameter from θ_{pr} , where θ_{mult} now measures the fraction of the people who would vote for the candidate of interest among those who would still vote but for somebody else without the persuasive treatment. The advantage of having a multinomial model is that we can pay attention to this extra layer of conditioning that comes from the presence of an outside option.

Similarly to the binary case, Assumption B1 enables us to express θ_{mult} in terms of the marginal distributions of the potential outcomes.

Lemma B1. *Under Assumption B1, we have*

$$\theta_{\text{mult}} = \frac{\mathbb{P}\{Y_{i1}(1) = 1\} - \mathbb{P}\{Y_{i1}(0) = 1\}}{1 - \mathbb{P}\{Y_{i0}(0) = 1\} - \mathbb{P}\{Y_{i1}(0) = 1\}}.$$

Lemma B1 shows that the conditional probability θ_{mult} can be obtained by simply rescaling the ATE as before. One complication (compared with the binary case) is that we have an extra term in the denominator (i.e., $\mathbb{P}\{Y_{i0}(0) = 1\}$), which substantially complicates the identification analysis.

The assumption of $Y_{i1}(1) \geq Y_{i1}(0)$ is sufficient for Lemma B1, but it is not necessary. Indeed, Assumption B1 rules out the possibility of having $Y_i(1) = (1, 0, 0)$ and $Y_i(0) = (0, 1, 0)$, but this is an irrelevant event for θ_{mult} because θ_{mult} focuses only on the case where $Y_{i0}(t) = 0$ for both $t = 0, 1$. However, Assumption B1 turns out to be quite convenient to obtain informative bounds of θ_{mult} .

¹⁷Here, we note that there is no loss of generality in assuming that \mathcal{S} has only three options; if not, we can simply define $Y_{i,-1}(t) = \sum_{j \in \mathcal{S} \setminus \{0, 1\}} Y_{ij}(t)$.

Assumption B2 (No Defiers and an Exogenous Instrument). T_i satisfies Assumption B. Further, for $t = 0, 1$ and $j \in \{0, 1, -1\}$, $(Y_{ij}(t), V_i)$ is independent of Z_i .

Assumption B2 is a trivial extension of Assumption B. In the following subsection, we provide a complete characterization of the sharp identified set of θ_{mult} under Assumptions B1 and B2 in each of the three scenarios of data availability. However, we present here only the sharp lower bound on θ_{mult} when the joint distribution of (Y_i, T_i, Z_i) is available, as it seems to be the most useful result in practice. Here, $Y_i = (Y_{i0}, Y_{i1}, Y_{i,-1}) = T_i Y_i(1) + (1 - T_i) Y_i(0)$ is now a three-dimensional vector of observed binary variables.

Theorem B1. Suppose that Assumptions B1 and B2 are satisfied and that the joint distribution of (Y_i, T_i, Z_i) is known, where $\mathbb{P}(Y_{i,-1} = 1 \mid Z_i = 0) > 0$. If $\mathbb{P}(Y_{i1} = 1 \mid Z_i = 1) + \mathbb{P}(Y_{i0} = 1, T_i = 0 \mid Z_i = 0) < 1$, then the lower bound of the sharp identified interval of θ_{mult} is given by

$$\theta_{L,\text{mult}} := \frac{\mathbb{P}(Y_{i1} = 1 \mid Z_i = 1) - \mathbb{P}(Y_{i1} = 1 \mid Z_i = 0)}{1 - \mathbb{P}(Y_{i0} = 1, T_i = 0 \mid Z_i = 0) - \mathbb{P}(Y_{i1} = 1 \mid Z_i = 0)}.$$

Theorem B1 shows the sharp lower bound of θ_{mult} in the most favorable data scenario; the complete characterization of the sharp identified set of θ_{mult} in the other data scenarios can be found in online Appendix B.2. The condition of $\mathbb{P}(Y_{i1} = 1 \mid Z_i = 1) + \mathbb{P}(Y_{i0} = 1, T_i = 0 \mid Z_i = 0) \geq 1$ represents an extreme situation, in which case θ_{mult} can be shown to be equal to 1. The intuition is clear; if there are too many people who do not vote while they are untreated, then there are too few people to “persuade” as θ_{mult} focuses only on the group of people who would vote even without the treatment. Including this trivial situation, the sharp lower bound $\theta_{L,\text{mult}}$ is always no smaller than that of the binary persuasion rate. Therefore, ignoring the presence of an outside option can lead to underestimating the persuasive effect of the treatment if θ_{mult} is the parameter of interest. In fact, in the empirical example in Section 6, we estimate the lower bound of θ_{mult} in Equation (B1) with the outside action of not voting at all. We find that the estimated lower bound on the average persuasion increases from 0.0707 to 0.0975 if we condition on those who would vote without reading the newspaper.

Unlike the lower bound θ_L in the binary case, $\theta_{L,\text{mult}}$ depends on the joint distribution of (Y_i, T_i) given Z_i . Therefore, if the sampling scheme does not allow the econometrician to access the full joint distribution, then even the lower bound will change. By applying a version of the Fréchet–Hoeffding bounds, we can derive the sharp identified bounds of θ_{mult} under the other two sampling schemes (i.e., Assumptions D and E). Not surprisingly, in the least informative case of Assumption E, the sharp lower bound of θ_{mult} becomes identical to that of θ_{pr} (i.e., θ_L).

B.2. Complete Characterization of the Identified Sets. In this section we present a complete characterization of the sharp identified set with nonbinary outcomes under the three scenarios of data availability.

In order to present the results, we introduce the following notation. Let

$$p_j(y | z) := \mathbb{P}(Y_{ij} = y | Z_i = z) \quad \text{and} \quad p_j(y, t | z) := \mathbb{P}(Y_{ij} = y, T_i = t | Z_i = z). \quad (\text{B2})$$

The first theorem is a complete characterization of the sharp identified set of θ_{mult} when (Y_i, T_i, Z_i) is jointly observed.

Theorem B2. *Suppose that Assumptions B1 and B2 are satisfied and that the joint distribution of (Y_i, T_i, Z_i) is known, where $p_{-1}(1 | 0) > 0$. Then, the sharp identified set of θ_{mult} is given as follows.*

- (i) If $p_1(1 | 1) + p_0(1, 0 | 0) \geq 1$, then $\theta_{\text{mult}} = 1$;
- (ii) If $p_1(1 | 1) + p_0(1, 0 | 0) < 1 \leq p_1(1, 1 | 1) + p_0(1, 0 | 0) + 1 - e(1) + e(0)$, then

$$\frac{p_1(1 | 1) - p_1(1 | 0)}{1 - p_0(1, 0 | 0) - p_1(1 | 0)} \leq \theta_{\text{mult}} \leq 1;$$

- (iii) If $p_1(1, 1 | 1) + p_0(1, 0 | 0) + 1 - e(1) + e(0) < 1$, then

$$\frac{p_1(1 | 1) - p_1(1 | 0)}{1 - p_0(1, 0 | 0) - p_1(1 | 0)} \leq \theta_{\text{mult}} \leq \frac{p_1(1, 1 | 1) + 1 - e(1) - p_1(1, 0 | 0)}{1 - p_0(1, 0 | 0) - e(0) - p_1(1, 0 | 0)}.$$

We now present the results under Assumption D.

Theorem B3. *Suppose that Assumptions B1 and B2 are satisfied and that the distribution of (Y_i, Z_i) and $e(1), e(0)$ are known, where $p_{-1}(1 | 0) > 0$. Then, the sharp identified set of θ_{mult} is given as follows.*

- (i) If $p_1(1 | 1) + \max\{0, p_0(1 | 0) - e(0)\} \geq 1$, then $\theta_{\text{mult}} = 1$;
- (ii) If $p_1(1 | 1) + \max\{0, p_0(1 | 0) - e(0)\} < 1 \leq \min\{p_1(1 | 1), e(1)\} + \min\{p_0(1 | 0), 1 - e(0)\} + 1 - e(1) + e(0)$, then

$$\max \left\{ \frac{p_1(1 | 1) - p_1(1 | 0)}{1 - p_1(1 | 0)}, \frac{p_1(1 | 1) - p_1(1 | 0)}{1 - p_0(1 | 0) + e(0) - p_1(1 | 0)} \right\} \leq \theta_{\text{mult}} \leq 1;$$

- (iii) If $\min\{p_1(1 | 1), e(1)\} + \min\{p_0(1 | 0), 1 - e(0)\} + 1 - e(1) + e(0) < 1$, then

$$\begin{aligned} \max \left\{ \frac{p_1(1 | 1) - p_1(1 | 0)}{1 - p_1(1 | 0)}, \frac{p_1(1 | 1) - p_1(1 | 0)}{1 - p_0(1 | 0) + e(0) - p_1(1 | 0)} \right\} &\leq \theta_{\text{mult}} \\ &\leq \frac{\min\{p_1(1 | 1), e(1)\} + 1 - e(1) - \max\{0, p_1(1 | 0) - e(0)\}}{1 - \min\{p_0(1 | 0), 1 - e(0)\} - e(0) - \max\{0, p_1(1 | 0) - e(0)\}}. \end{aligned}$$

The three cases in Theorem B3 are exclusive and exhaustive. The condition describing the third case can be simplified to $p_1(1, 1 | 1) + p_0(1 | 0) < e(1) - e(0)$ because the inequality can hold only if $\min\{p_1(1 | 1), e(1)\} = p_1(1 | 1)$ and $\min\{p_0(1 | 0), 1 - e(0)\} = p_0(1 | 0)$. We finally give the result under the assumption that T_i is not observed at all.

Theorem B4. *Suppose that Assumptions B1 and B2 are satisfied and that only the distribution of (Y_i, Z_i) is known, where $p_{-1}(1 | 0) > 0$. Then, the sharp identified set of θ_{mult} is the same as the binary case, i.e.,*

$$\frac{p_1(1 | 1) - p_1(1 | 0)}{1 - p_1(1 | 0)} \leq \theta_{\text{mult}} \leq 1,$$

provided that the denominator is different from zero.

Therefore, unlike the binary case, the sharp lower bound on θ_{mult} depends on the sampling scheme. If T_i is not observed at all, then observing the choices of the outside option does not help to infer the persuasion rate with the extra conditioning on those who would participate without the treatment.

PART II. ADDITIONAL EMPIRICAL RESULTS

In this part of the appendices, we provide additional empirical results. Before going through details, Table B1 provides a concise summary of additional empirical results that report either interval or point estimates of both the average and local persuasion rates (θ_{avg} and θ_{local}). Table B1 shows that the persuasive effect is highly heterogeneous.

APPENDIX C. FURTHER DISCUSSION ON SECTION 6

C.1. Opting into the Free Subscription to The Washington Post. In Section 6, we provide an empirical example using the data from Gerber, Karlan, and Bergan (2009). As we describe in the main text, the main difference between our example and that in DellaVigna and Gentzkow (2010) is what persuasive treatment is concerned with. In the latter, binary treatment is $T_i = 1$ if the i th individual opted into the free subscription and $T_i = 0$ if they opted out of it. We believe DellaVigna and Gentzkow (2010)'s analysis is closer to an intent-to-treat (ITT) analysis because merely opting into the free subscription does not mean that individuals were exposed to the persuasive messages of the newspaper unless they actually read them. Furthermore, the difficulty of using the treatment variable as in DellaVigna and Gentzkow (2010) is that there is no individual-level information regarding who opted out of the free subscription from the original dataset of Gerber, Karlan, and Bergan (2009). DellaVigna and Gentzkow (2010) computed the changes in the exposure rate $e(1) - e(0) = 94\%$ from Gerber, Karlan, and Bergan (2009, page 38), which states that "6 percent of households in the treatment groups opted out of the free subscription." Given the lack of micro-level information, we can only compute the bounds using summary statistics on $\mathbb{P}(Y_i = 1 | Z_i = 1)$, $\mathbb{P}(Y_i = 1 | Z_i = 0)$, $e(1) = \mathbb{P}(T_i = 1 | Z_i = 1)$, and $e(0) = \mathbb{P}(T_i = 1 | Z_i = 0)$. This case corresponds to Theorem 4 in the paper. It follows from Table 1C of Gerber, Karlan, and Bergan (2009) that 49% and 41% voted for a Democrat in 2005 VA election in the Washington Post treatment and control group, respectively: i.e., $\hat{\mathbb{P}}(Y_i = 1 | Z_i = 1) = 0.49$ and $\hat{\mathbb{P}}(Y_i = 1 | Z_i = 0) = 0.41$. Since the control group in Gerber, Karlan, and Bergan (2009) did not receive an offer to the free subscription, we have $\hat{e}(1) = 0.94$ and $\hat{e}(0) = 0$. These numbers result in

TABLE B1. Summary of Additional Empirical Results: Estimates of Persuasion Rates

Treatment	θ_{avg}	θ_{local}
DellaVigna and Kaplan (2007)		
Watching Fox News	[0.005, 0.997]	[0.117, 1]
Watching Fox News	[0.011, 0.997]	[0.375, 1]
Landry, Lange, List, Price, and Rupp (2006)		
VCM	[0.095, 0.719]	0.253
VCM with seed money	[0.052, 0.699]	0.148
Single-prize lottery	[0.171, 0.794]	0.455
Multiple-prize lottery	[0.126, 0.775]	0.359
DellaVigna, List, and Malmendier (2012)		
La Rabida Children's Hospital		
Baseline	[0.071, 0.666]	0.175
Flyer	[0.068, 0.704]	0.188
Opt-Out	[0.054, 0.749]	0.177
DellaVigna, List, and Malmendier (2012)		
East Carolina Hazard Center		
Baseline	[0.047, 0.617]	0.109
Flyer	[0.051, 0.655]	0.129
Opt-Out	[0.030, 0.686]	0.086

^a The first row for **DellaVigna and Kaplan (2007)** refers to estimates when U.S. House district fixed effects are controlled for; the second row corresponds to those with county fixed effects.

^b VCM stands for voluntary contributions mechanism.

the bounds $[0.136, 0.237]$ for the average persuasion rate θ_{avg} and $[0.136, 1]$ for the local persuasion rate θ_{local} , respectively. On the other hand, the estimate of $\tilde{\theta}_{DK}$ is 0.144. The difference between θ_L and $\tilde{\theta}_{DK}$ is smaller here because the size of compliers is large (that is, 94%) when treatment refers to opting into the free subscription.

APPENDIX D. ADDITIONAL EXAMPLES: THE EFFECTS OF MEDIA ON VOTING

In this section we revisit the recent empirical literature on the effects of media on voting and apply our identification results.

D.1. The Effect of Fox News: DellaVigna and Kaplan (2007) Revisited. In DK, the entry of Fox News in cable markets plays a role of an instrument conditional on a set of covariates. That is, Z_i is a binary variable that equals one if Fox News was a part of the local cable package in the town where the i^{th} individual was living in 2000. To apply our result to DK, let Y_i be the binary dependent variable that equals one if individual i voted for the Republican candidate in the 2000 presidential election. As DK argue in their paper, Fox News availability in 2000 is likely to be idiosyncratic, only after controlling for a set of covariates. We will be explicit about conditioning on covariates X_i to apply our identification results: i.e., using Bayes' theorem,

$$\begin{aligned}\theta_{\text{avg}} &= \int \frac{\mathbb{P}\{Y_i(1) = 1 \mid X_i = x\} - \mathbb{P}\{Y_i(0) = 1 \mid X_i = x\}}{\mathbb{P}\{Y_i(0) = 0 \mid X_i = x\}} dF\{x \mid Y_i(0) = 0\} \\ &= \frac{\int \mathbb{P}\{Y_i(1) = 1 \mid X_i = x\} dF(x) - \int \mathbb{P}\{Y_i(0) = 1 \mid X_i = x\} dF(x)}{1 - \int \mathbb{P}\{Y_i(0) = 1 \mid X_i = x\} dF(x)},\end{aligned}$$

where F denote the marginal distribution of X_i . Since the sharp identified bounds of $\mathbb{P}\{Y_i(t) = 1 \mid X_i = x\}$ for $t = 0, 1$ can be obtained in the same way as lemmas I5 and I6, depending on the data availability scenario, the same reasoning as the proof of Theorems 2, 4 and 5 leads to the sharp bounds on θ_{avg} when we have X_i in the data.

To be more specific, let $e(z, x) = \mathbb{P}(T_i = 1 \mid Z_i = z, X_i = x)$, and we focus on the second data scenario, i.e., the case of Assumption D, given the type of DK's data. Then, similarly to Lemma I6, the sharp identified bounds of $\mathbb{P}\{Y_i(t) = 1 \mid X_i\}$ are given by

$$\begin{aligned}\mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i) &\leq \mathbb{P}\{Y_i(1) = 1 \mid X_i\} \leq \min\{1, \mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i) + 1 - e(1, X_i)\}, \\ \max\{0, \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i) - e(0, X_i)\} &\leq \mathbb{P}\{Y_i(0) = 1 \mid X_i\} \leq \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i),\end{aligned}$$

from which

$$\begin{aligned}\int \mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x) dF(x) \\ \leq \mathbb{P}\{Y_i(1) = 1\} \leq \int \min\{1, \mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x) + 1 - e(1, x)\} dF(x),\end{aligned}$$

and

$$\begin{aligned}\int \max\{0, \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x) - e(0, x)\} dF(x) \\ \leq \mathbb{P}\{Y_i(0) = 1\} \leq \int \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x) dF(x).\end{aligned}$$

Therefore, following the proof of theorem 2 shows that the sharp identified bounds of θ_{avg} when X_i is in the data can be obtained by integrating both the numerators and the denominators of the "conditional" sharp identified bounds with respect to the marginal distribution of X_i .

In order to implement this idea, we write the lower bound as a function of the values of X_i : i.e.,

$$\tilde{\theta}_L(x) := \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x) - \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x)}{1 - \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x)}, \quad (\text{D1})$$

which is the sharp lower bound on the conditional persuasion rate, i.e., $\mathbb{P}\{Y_i(1) = 1 \mid Y_i(0) = 0, X_i = x\}$. Then, to obtain the lower bound for the persuasion rate in the population, we integrate the numerator and the denominator of (D1) with respect to the distribution F of X_i , so that

$$\theta_L = \frac{\int \mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x) dF(x) - \int \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x) dF(x)}{1 - \int \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x) dF(x)}. \quad (\text{D2})$$

Note that X_i is first controlled for and is averaged out.

To estimate θ_L in this way by using DK's data,¹⁸ we adopt similar specifications as in DK. They estimated $\mathbb{P}(Y_i = 1 \mid Z_i, X_i)$ using a town-level linear regression model, where the dependent variable is the Republican two-party vote share for the 2000 presidential election minus the same variable for the 1996 election. To be consistent with our econometric framework, we modify the dependent variable to be the votes cast for the Republican candidate in the 2000 presidential election divided by the population of age 18 and older. Recall that in our setup, $Y_i = 0$ if individual i did not vote for the Republican candidate. This event includes the case of voting for different candidates or that of not voting for any candidate at all. As for the town-level covariates, we include the Republican vote share as a share of the voting-age population in the 1996 election, census controls for both 1990 and 2000, cable system controls, and U.S. House district fixed effects (or county fixed effects). These specifications correspond to the main specifications of DK (see columns (4) and (5) of table IV in DK). In the regression, the town-level observations are weighted by the population of age 18 and older in 1996.

DK used two different data sources for (Y_i, Z_i, X_i) and (T_i, Z_i, X_i) . Hence, we can look at the upper bound for θ_{avg} and the lower bound for θ_{local} using these. Again, making use of the covariates explicitly, we use the conditional version of lemma I6 and follow the same reasoning as the proof of Theorem 4 to obtain

$$\theta_{U_e} = \frac{\int \tilde{\theta}_{U_e,n}(x) dF(x)}{\int \tilde{\theta}_{U_e,d}(x) dF(x)},$$

where

$$\begin{aligned} \tilde{\theta}_{U_e,n}(X_i) &:= \min\{1, \mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i) + 1 - e(1, X_i)\} - \max\{0, \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i) - e(0, X_i)\}, \\ \tilde{\theta}_{U_e,d}(X_i) &:= 1 - \max\{0, \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i) - e(0, X_i)\}. \end{aligned}$$

The case of θ_{local} when we have X_i in the data is similar. Let $\mathcal{C}_i = \{e(0, X_i) < V_i \leq e(1, X_i)\}$ denote the event of individual i being a complier. We note that

$$\theta_{\text{local}} = \int \mathbb{P}\{Y_i(1) = 1 \mid Y_i(0) = 0, X_i = x, \mathcal{C}_i\} dF(x \mid Y_i(0) = 0, \mathcal{C}_i)$$

¹⁸The data used in DK are available at <http://eml.berkeley.edu/~sdellavi/index.html>.

$$\begin{aligned}
&= \int \left(\frac{\mathbb{P}\{Y_i(1) = 1 \mid X_i = x, \mathcal{C}_i\} - \mathbb{P}\{Y_i(0) = 1 \mid X_i = x, \mathcal{C}_i\}}{\mathbb{P}\{Y_i(0) = 0 \mid X_i = x, \mathcal{C}_i\}} \right) dF(x \mid Y_i(0) = 0, \mathcal{C}_i) \\
&= \frac{\int \mathbb{P}\{Y_i(1) = 1 \mid X_i = x, \mathcal{C}_i\} - \mathbb{P}\{Y_i(0) = 1 \mid X_i = x, \mathcal{C}_i\} dF(x \mid \mathcal{C}_i)}{\mathbb{P}\{Y_i(0) = 0 \mid \mathcal{C}_i\}}
\end{aligned} \tag{D3}$$

by Bayes' theorem. Here, the numerator of (D3) is equal to

$$\int \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x) - \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x)}{e(1, x) - e(0, x)} dF(x \mid \mathcal{C}_i), \tag{D4}$$

and the denominator of (D3) can be rewritten as

$$\mathbb{P}\{Y_i(0) = 0 \mid \mathcal{C}_i\} = \int \mathbb{P}\{Y_i(0) = 0 \mid X_i = x, \mathcal{C}_i\} dF(x \mid \mathcal{C}_i). \tag{D5}$$

Since

$$dF(x \mid \mathcal{C}_i) = \frac{e(1, x) - e(0, x)}{\int \{e(1, x) - e(0, x)\} dF(x)} dF(x).$$

by Bayes' theorem again, we can combine (D3) to (D5) to obtain

$$\theta_{\text{local}} = \frac{\int \{\mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x) - \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x)\} dF(x)}{\int \mathbb{P}\{Y_i(0) = 0, e(0, x) < V_i \leq e(1, x) \mid X_i = x\} dF(x)} \tag{D6}$$

The numerator of (D6) is directly identified from data in all of the three data scenarios we discussed.

The denominator of (D6) is also identified under Assumption C by

$$\int \mathbb{P}(Y_i = 0, T_i = 0 \mid Z_i = 0, X_i = x) - \mathbb{P}(Y_i = 0, T_i = 0 \mid Z_i = 1, X_i = x) dF(x),$$

but it is not under Assumption D. The sharp identified bounds on the denominator of (D6) under Assumption D can be obtained exactly in the same reasoning as in the proof of part (ii) of Theorem 6.

Specifically, the sharp bounds on θ_{local} are now given by

$$\frac{\int \{\mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x) - \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x)\} dF(x)}{\int \min\{1 - \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x), e(1, x) - e(0, x)\} dF(x)} \leq \theta_{\text{local}} \leq 1 \tag{D7}$$

under Assumption D. The lower bound in (D7) is always no smaller than

$$\max \left\{ \frac{\int \{\mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x) - \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x)\} dF(x)}{1 - \int \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x) dF(x)}, \frac{\int \{\mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x) - \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x)\} dF(x)}{\int \{e(1, x) - e(0, x)\} dF(x)} \right\},$$

where the first expression in the max function is the sharp lower bound θ_L given in (D2) and the second one is the average of the conditional local average treatment effect over the group of compliers.

DK estimated $e(z, x)$ using the microlevel Scarborough data on television audiences. We focus on “diary audience” measure in DK¹⁹ and take the same specifications as in columns (2) and (3) of table VIII from DK.

TABLE D1. Persuasion Rates: Fox News Effects

	(1)	(2)
	U.S. House district	County
	fixed effects	fixed effects
θ_{avg}	[0.005,0.997]	[0.011,0.997]
θ_{local}	[0.117,1]	[0.375,1]

Table D1 summarizes our empirical results.²⁰ Column (1) shows estimation results when U.S. House district fixed effects are controlled for and column (2) displays corresponding results for county fixed effects.

The bounds for θ_{avg} are wide and uninformative. However, the lower bounds for θ_{local} are sizable and also comparable to the estimates of the persuasion rates reported in DK (0.11 and 0.28, respectively). In sum, we conclude that the persuasive effect of Fox News seems fairly large for the compliers, that is, those who would watch the Fox News channel if and only if it is randomly available, although the data do not say much about the entire population.

D.2. The NTV Effect: Enikolopov, Petrova, and Zhuravskaya (2011) Revisited. As mentioned in the main text, Enikolopov, Petrova, and Zhuravskaya (2011, EPZ hereafter) used a continuous instrument, i.e., the signal strength of NTV, to measure the persuasive effect of watching NTV (the anti-Putin TV station) on a parliamentary election in 1999. Further, in the individual-level survey data in EPZ, (Y_i, T_i, Z_i) are jointly observed. Therefore, in this subsection, we apply the identification result of the marginal persuasion rate to this example using the EPZ data.

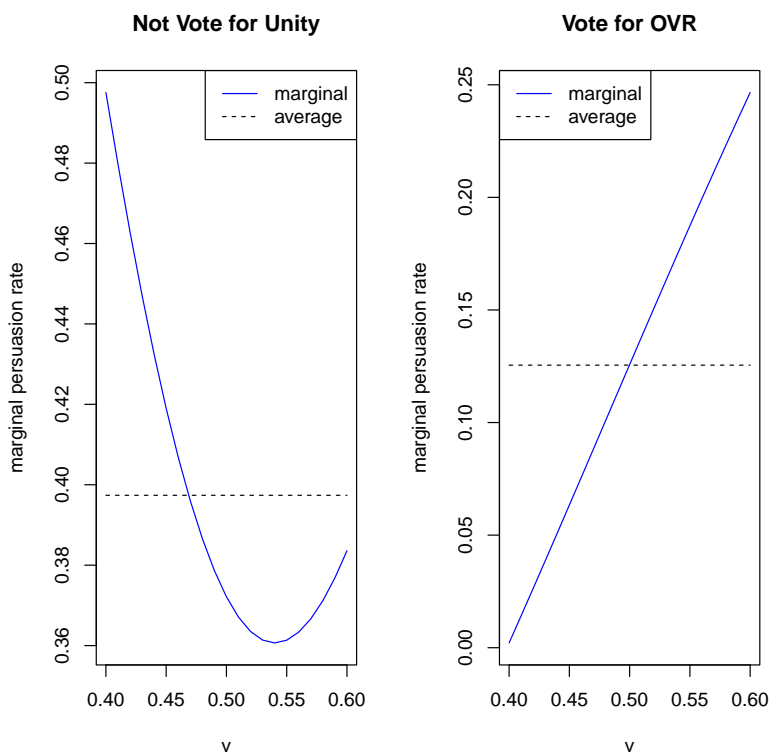
We look at two parties: the progovernment party “Unity” and the most popular opposition party OVR (“Fatherland–All Russia”). During the 1999 election campaign, Unity was opposed by NTV, while OVR were supported by NTV. Thus, EPZ presumed a negative persuasion rate for voting for Unity but a positive persuasion rate for OVR. To be consistent with our theoretical framework and other empirical examples, Y_i is either $Y_{\text{Unity},i}$ or $Y_{\text{OVR},i}$, depending on which party we consider. Specifically, we let $Y_{\text{Unity},i} = 1$ if an individual did not vote for Unity and $Y_{\text{Unity},i} = 0$ otherwise; $Y_{\text{OVR},i} = 1$ if an individual voted for OVR and $Y_{\text{OVR},i} = 0$ otherwise. As in the previous section, it is necessary to condition on covariates. We take the baseline covariates as in columns (1) and (2) of tables 6 and 7 in EPZ. They

¹⁹The microlevel Scarborough data contain the “recall” measure regarding whether a respondent watched a given channel in the past seven days and the “diary” measure on whether a respondent watched a channel for at least one full half-an-hour block according to the seven-day diary.

²⁰To estimate the unconditional bounds reported in the table, the conditional ones are weighted by the number of respondents in a town for the Scarborough data. In addition, the predicted probabilities are truncated to be between 0 and 1.

include individual demographic characteristics such as gender, age, marital status, and education, and subregional variables such as population size and average wage.

FIGURE D1. Estimates of Marginal and Average Persuasion Rates



Notes: The left and right panels of the figure show estimates of the marginal and average persuasion rates for not voting for Unity and voting for OVR, respectively. The domain of V_i is limited to $[0.4, 0.6]$ and the average persuasion rates here refer to the averages of the marginal persuasion rates between 0.4 and 0.6.

Since we have covariates X_i , $\theta_{\text{marginal}}(v)$ is equal to

$$\begin{aligned}
 & \int \mathbb{P}\{Y_i(1) = 1 \mid Y_i(0) = 0, V_i = v, X_i = x\} dF\{x \mid Y_i(0) = 0, V_i = v\} \\
 &= \int \frac{\mathbb{P}\{Y_i(1) = 1 \mid V_i = v, X_i = x\} - \mathbb{P}\{Y_i(0) = 1 \mid V_i = v, X_i = x\}}{\mathbb{P}\{Y_i(0) = 0 \mid V_i = v, X_i = x\}} dF\{x \mid Y_i(0) = 0, V_i = v\} \\
 &= \frac{\int \mathbb{P}\{Y_i(1) = 1 \mid V_i = v, X_i = x\} - \mathbb{P}\{Y_i(0) = 1 \mid V_i = v, X_i = x\} dF(x)}{1 - \int \mathbb{P}\{Y_i(0) = 1 \mid V_i = v, X_i = x\} dF(x)},
 \end{aligned}$$

where the last equality is by Bayes' theorem and the independence between V_i and X_i . Therefore, $\theta_{\text{marginal}}(v)$ is obtained by integrating both the numerator and the denominator of the conditional marginal persuasion rate with respect to the marginal distribution of X_i : i.e.,

$$\theta_{\text{marginal}}(v) = \frac{\int \partial \mathbb{P}\{Y_i = 1 \mid e(Z_i, X_i) = e, X_i = x\} / \partial e|_{e=v} dF(x)}{1 + \int \partial \mathbb{P}\{Y_i = 1, T_i = 0 \mid e(Z_i, X_i) = e, X_i = x\} / \partial e|_{e=v} dF(x)}. \quad (\text{D8})$$

For the sake of simplicity, we estimate $\theta_{\text{marginal}}(v)$ via polynomial sieves: the relevant conditional probabilities, i.e., $e(z, x)$, $\mathbb{P}(Y_i = 1 \mid e(Z_i, X_i) = e, X_i = x)$ and $\mathbb{P}(Y_i = 1, T_i = 0 \mid e(Z_i, X_i) = e, X_i = x)$, are all estimated by regression models that are linear in X_i and cubic in either Z_i or $e(Z_i, X_i)$, depending on the conditional probability to be estimated, while interacting demographic variables in X_i with Z_i or $e(Z_i, X_i)$, respectively.²¹ Then, we obtain the conditional marginal persuasion rate $\theta_{\text{marginal}}(v, x)$, i.e., the estimates of Equation (12) with extra conditioning on $X_i = x$. Finally, we integrate x out of the numerator and the denominator of $\theta_{\text{marginal}}(v, x)$ by using the sample marginal distribution of X_i as we illustrated in (D8).

Figure D1 presents the estimation results. As it is more difficult to estimate $\theta_{\text{marginal}}(v)$ for high or low values of v , we focus on the middle range of v , i.e., $v \in [0.4, 0.6]$: the average persuasion rate on the middle range of v is computed by

$$\theta_{\text{avg}}[0.4, 0.6] = \frac{\int_{0.4}^{0.6} \int \partial \mathbb{P}\{Y_i = 1 \mid e(Z_i, X_i) = e, X_i = x\} / \partial e|_{e=v} dF(x) dv}{\int_{0.4}^{0.6} \left[1 + \int \partial \mathbb{P}\{Y_i = 1, T_i = 0 \mid e(Z_i, X_i) = e, X_i = x\} / \partial e|_{e=v} dF(x) \right] dv}.$$

In the left panel, $\theta_{\text{marginal}}(v)$ and θ_{avg} are plotted as a function of v , when the outcome variable is not to vote for Unity. It can be seen that the marginal persuasive rate is about 50% at $v = 0.4$ but just 40% for $v > 0.5$. In view of Equation (5), V_i can be interpreted as the unobserved cost of watching NTV. The estimation results suggest that the persuasive effect for not voting for Unity is stronger for those whose unobserved cost of watching NTV is lower. In the right panel, corresponding results are shown for OVR. In this case, the persuasive effect is weaker for those with lower values of v .

A striking pattern we can learn from Figure D1 is that persuasive effects are highly heterogeneous. This may partially answer the puzzle reported in EPZ. They found relatively modest positive persuasive effects for opposition parties but much stronger persuasive effects for Unity using aggregate voting outcomes, while the magnitudes are similar using individual survey data.²² Our estimation results indicate that the marginal persuasive effects are highly heterogeneous, thereby implying that different aggregate averages can be substantially different from each other. The average persuasive effect $\theta_{\text{avg}}[0.4, 0.6]$ is plotted as a horizontal line in each panel of Figure D1: it is 39.7% against Unity and 12.5% for OVR. In short, this application exemplifies the identification power of continuous instruments that can uncover the patterns of heterogeneity in persuasive effects.

APPENDIX E. ADDITIONAL EXAMPLES: DOOR-TO-DOOR FUNDRAISING

Landry, Lange, List, Price, and Rupp (2006) and DellaVigna, List, and Malmendier (2012) designed field experiments of door-to-door fundraising to examine various aspects of charity giving. In this section, we use their data to illustrate the usefulness of our identification results.

²¹The exposure rate $e(z, x)$ is first estimated and its fitted values, truncated to be between 0 and 1, are included as a regressor to estimate the other two conditional probabilities.

²²EPZ estimated the persuasion rate using a continuous version of DK. See equations (3) and (4) in EPZ for their formulae.

The common data structure in both papers is that for each type of experimental treatments, we observe (Y_i, T_i, Z_i) :

- $Y_i = 1$ if a household made a contribution to door-to-door fundraising,
- $T_i = 1$ if a household answered the door and spoke to a solicitor,
- $Z_i = 1$ if a household was approached by a solicitor.

If $Z_i = 0$ (a household was not approached by a solicitor), then $T_i = 0$ and furthermore it is very likely that $Y_i = 0$. Hence, in this section, we assume that $\mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) = \mathbb{P}(Y_i = 1 \mid Z_i = 0) = 0$. In addition, we assume that if $Y_i = 1$, it must be the case that $T_i = 1$. In other words, we assume that it is impossible to have both $Y_i = 1$ and $T_i = 0$ (making a contribution without answering the door). Thus, $\mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) = \mathbb{P}(Y_i = 1 \mid Z_i = 1)$. These assumptions were also used in computation of the persuasion rates for donors in [DellaVigna and Gentzkow \(2010\)](#). Under these assumptions, we have the bound for θ_{avg} as

$$\theta_L = \mathbb{P}(Y_i = 1 \mid Z_i = 1) \quad \text{and} \quad \theta_U = \mathbb{P}(Y_i = 1 \mid Z_i = 1) + 1 - e(1).$$

In addition,

$$\theta_{\text{local}} = \mathbb{P}(Y_i = 1 \mid Z_i = 1) / e(1);$$

θ_{local} is the same as the usual LATE.

E.1. Landry, Lange, List, Price, and Rupp (2006) Revisited. In this study, there were four treatments: VCM (voluntary contributions mechanism), VCM with seed money, single-prize lottery, and multiple-prize lottery. Using Table II of [Landry, Lange, List, Price, and Rupp \(2006\)](#), we compute the persuasive effects by treatment and report results in Table E1.

TABLE E1. Persuasive Effect by Treatment in [Landry, Lange, List, Price, and Rupp \(2006\)](#)

Treatment	$\mathbb{P}(Y_i = 1 \mid Z_i = 1)$	$e(1)$	θ_L	θ_U	θ_{local}
VCM	9.5%	37.6%	9.5%	71.9%	25.3%
VCM with seed money	5.2%	35.3%	5.2%	69.9%	14.8%
Single-prize lottery	17.1%	37.7%	17.1%	79.4%	45.5%
Multiple-prize lottery	12.6%	35.2%	12.6%	77.5%	35.9%
All	10.8%	36.3%	10.8%	74.5%	29.7%

Based on the lower bound and the LATE parameter, it seems that the single-prize lottery is the most effective fundraising tool, whereas the VCM with seed money is the least effective. However, the identification regions for θ_{avg} of all four treatments overlap and there is no clear ranking based on those. This suggests that if one cares about the persuasive effect for the population, the evidence is inconclusive.

E.2. **DellaVigna, List, and Malmendier (2012) Revisited.** In their study of charity giving, DellaVigna, List, and Malmendier (2012, DLM hereafter) designed both fundraising and survey treatments to test for altruism and social pressure in charity giving. In this section, we focus only on three fundraising treatments: namely, the baseline treatment, the flyer treatment, and the opt-out treatment. The baseline treatment is the standard door-to-door funding raising campaign, the flyer treatment is with the flyer that provided information on fundraising the date before the solicitation, and the opt-out treatment is with the flyer that had an additional feature of a “Do Not Disturb” checkbox. There were two charities in each of the fundraising treatments: La Rabida Children’s Hospital and the East Carolina Hazard Center.

TABLE E2. Persuasive Effect by Treatment in DLM

Treatment	$\mathbb{P}(Y_i = 1 \mid Z_i = 1)$	$e(1)$	θ_L	θ_U	θ_{local}
La Rabida Children’s Hospital					
Baseline	7.1%	40.5%	7.1%	66.6%	17.5%
Flyer	6.8%	36.4%	6.8%	70.4%	18.8%
Opt-Out	5.4%	30.4%	5.4%	74.9%	17.7%
East Carolina Hazard Center					
Baseline	4.7%	43.0%	4.7%	61.7%	10.9%
Flyer	5.1%	39.6%	5.1%	65.5%	12.9%
Opt-Out	3.0%	34.4%	3.0%	68.6%	8.6%

DLM pointed out that treatments were randomized within a date-solicitor time block and estimated linear probability models with covariates: solicitor fixed effects, date-town fixed effects, hourly time block fixed effects, and area rating dummies. We use the same specification as in DLM, estimate $\mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x)$ and $e(1, x)$, and then average out the conditional estimates as in Appendix D.1.²³ The resulting estimates are reported in Table E2, where we report the persuasive effect by treatment/charity.

The local persuasion rate is point identified and is higher for the in-state charity, La Rabida Children’s Hospital. The estimates of θ_{local} are the highest for the flyer treatment in both charities. This does not mean that the flyer treatment is the most effective in fundraising for the general population. Note that the compliers of the baseline treatment are different from those of the flyer treatment. For example, it could be the case that households at the margin of giving might have decided not to answer the door after they noticed the flyer. Unlike θ_{local} , θ_L and θ_U are comparable across different treatments. However, as in the previous section, it is difficult to see whether there is a significant difference across treatments if we focus on the bounds for θ_{avg} .²⁴

²³The data collected in DLM are available at <http://eml.berkeley.edu/~sdellavi/index.html>. As before, the predicted probabilities are truncated to be between 0 and 1, when they are averaged out.

²⁴In addition to the fundraising treatments, DLM relied on survey treatments and structural estimates to draw conclusions in their paper.

PART III. INFERENCE AND SEMIPARAMETRIC ESTIMATION

APPENDIX F. INFERENCE ON THE AVERAGE PERSUASION RATE

In this section, we provide methods for carrying out inference on the average persuasion rate, for which we assume that the data are independent and identically distributed.

F.1. The case of Theorem 2. In this subsection, we consider Theorem 2, where the sharp identified interval of θ_{avg} is given by $[\theta_L, \theta_U]$. Recall that

$$\theta_L = \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}{1 - \mathbb{P}(Y_i = 1 \mid Z_i = 0)},$$

$$\theta_U = \frac{\mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) + 1 - e(1)}{1 - \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0)}.$$

To define the sample analog estimators of θ_L and θ_U , define

$$\hat{\mathbb{P}}(Y_i = 1 \mid Z_i = 1) = \frac{\sum_{i=1}^n \mathbb{1}(Y_i = 1, Z_i = 1)}{\sum_{i=1}^n \mathbb{1}(Z_i = 1)},$$

$$\hat{\mathbb{P}}(Y_i = 1 \mid Z_i = 0) = \frac{\sum_{i=1}^n \mathbb{1}(Y_i = 1, Z_i = 0)}{\sum_{i=1}^n \mathbb{1}(Z_i = 0)},$$

$$\hat{\mathbb{P}}(Y_i = 1, T_i = 1 \mid Z_i = 1) = \frac{\sum_{i=1}^n \mathbb{1}(Y_i = 1, T_i = 1, Z_i = 1)}{\sum_{i=1}^n \mathbb{1}(Z_i = 1)},$$

$$\hat{\mathbb{P}}(Y_i = 1, T_i = 0 \mid Z_i = 0) = \frac{\sum_{i=1}^n \mathbb{1}(Y_i = 1, T_i = 0, Z_i = 0)}{\sum_{i=1}^n \mathbb{1}(Z_i = 0)},$$

$$\hat{e}(1) = \frac{\sum_{i=1}^n \mathbb{1}(T_i = 1, Z_i = 1)}{\sum_{i=1}^n \mathbb{1}(Z_i = 1)}.$$

Then we define

$$\hat{\theta}_L = \frac{\hat{\mathbb{P}}(Y_i = 1 \mid Z_i = 1) - \hat{\mathbb{P}}(Y_i = 1 \mid Z_i = 0)}{1 - \hat{\mathbb{P}}(Y_i = 1 \mid Z_i = 0)},$$

$$\hat{\theta}_U = \frac{\hat{\mathbb{P}}(Y_i = 1, T_i = 1 \mid Z_i = 1) - \hat{\mathbb{P}}(Y_i = 1, T_i = 0 \mid Z_i = 0) + 1 - \hat{e}(1)}{1 - \hat{\mathbb{P}}(Y_i = 1, T_i = 0 \mid Z_i = 0)}.$$

Since $\hat{\theta}_L$ and $\hat{\theta}_U$ are asymptotically jointly normal, we follow [Imbens and Manski \(2004\)](#) and [Stoye \(2009\)](#) to construct a confidence interval for θ_{avg} . Standard arguments based on the delta method yield that $\hat{\theta}_L$ and $\hat{\theta}_U$ have asymptotically linear approximations:

$$\hat{\theta}_L - \theta_L = n^{-1} \sum_{i=1}^n \varphi_{L,i} + o_p(n^{-1/2}),$$

$$\hat{\theta}_U - \theta_U = n^{-1} \sum_{i=1}^n \varphi_{U,i} + o_p(n^{-1/2}),$$

where $\varphi_{L,i}$ and $\varphi_{U,i}$ are influence functions that can be approximated by the following sample analogs:

$$\hat{\varphi}_{L,i} = \frac{1}{[1 - \hat{\mathbb{P}}(Y_i = 1 \mid Z_i = 0)] \hat{\mathbb{P}}(Z_i = 1)} \left\{ \mathbb{1}(Y_i = 1, Z_i = 1) - \hat{\mathbb{P}}(Y_i = 1, Z_i = 1) \right\}$$

$$\begin{aligned}
& - \frac{1}{[1 - \widehat{\mathbb{P}}(Y_i = 1 \mid Z_i = 0)]} \frac{\widehat{\mathbb{P}}(Y_i = 1 \mid Z_i = 1)}{\widehat{\mathbb{P}}(Z_i = 1)} \left\{ \mathbb{1}(Z_i = 1) - \widehat{\mathbb{P}}(Z_i = 1) \right\} \\
& + \frac{\widehat{\theta}_L - 1}{[1 - \widehat{\mathbb{P}}(Y_i = 1 \mid Z_i = 0)]} \frac{1}{\widehat{\mathbb{P}}(Z_i = 0)} \left\{ \mathbb{1}(Y_i = 1, Z_i = 0) - \widehat{\mathbb{P}}(Y_i = 1, Z_i = 0) \right\} \\
& - \frac{\widehat{\theta}_L - 1}{[1 - \widehat{\mathbb{P}}(Y_i = 1 \mid Z_i = 0)]} \frac{\widehat{\mathbb{P}}(Y_i = 1 \mid Z_i = 0)}{\widehat{\mathbb{P}}(Z_i = 0)} \left\{ \mathbb{1}(Z_i = 0) - \widehat{\mathbb{P}}(Z_i = 0) \right\}, \\
\widehat{\varphi}_{U,i} &= \frac{1}{[1 - \widehat{\mathbb{P}}(Y_i = 1, T_i = 0 \mid Z_i = 0)]} \frac{1}{\widehat{\mathbb{P}}(Z_i = 1)} \\
& \times \left\{ \mathbb{1}(Y_i = 1, T_i = 1, Z_i = 1) - \widehat{\mathbb{P}}(Y_i = 1, T_i = 1, Z_i = 1) + \mathbb{1}(T_i = 0, Z_i = 1) - \widehat{\mathbb{P}}(T_i = 0, Z_i = 1) \right\} \\
& - \frac{1}{[1 - \widehat{\mathbb{P}}(Y_i = 1, T_i = 0 \mid Z_i = 0)]} \frac{\widehat{\mathbb{P}}(Y_i = 1, T_i = 1, Z_i = 1) + \widehat{\mathbb{P}}(T_i = 0, Z_i = 1)}{\widehat{\mathbb{P}}(Z_i = 1)} \\
& \times \left\{ \mathbb{1}(Z_i = 1) - \widehat{\mathbb{P}}(Z_i = 1) \right\} \\
& + \frac{\widehat{\theta}_U - 1}{[1 - \widehat{\mathbb{P}}(Y_i = 1, T_i = 0 \mid Z_i = 0)]} \frac{1}{\widehat{\mathbb{P}}(Z_i = 0)} \left\{ \mathbb{1}(Y_i = 1, T_i = 0, Z_i = 0) - \widehat{\mathbb{P}}(Y_i = 1, T_i = 0, Z_i = 0) \right\} \\
& - \frac{\widehat{\theta}_U - 1}{[1 - \widehat{\mathbb{P}}(Y_i = 1, T_i = 0 \mid Z_i = 0)]} \frac{\widehat{\mathbb{P}}(Y_i = 1, T_i = 0 \mid Z_i = 0)}{\widehat{\mathbb{P}}(Z_i = 0)} \left\{ \mathbb{1}(Z_i = 0) - \widehat{\mathbb{P}}(Z_i = 0) \right\}.
\end{aligned}$$

Now define

$$\widehat{\sigma}_L^2 = n^{-2} \sum_{i=1}^n \widehat{\varphi}_{L,i}^2, \quad \widehat{\sigma}_U^2 = n^{-2} \sum_{i=1}^n \widehat{\varphi}_{U,i}^2, \quad \text{and} \quad \widehat{\Delta} = \widehat{\theta}_U - \widehat{\theta}_L.$$

That is, $\widehat{\sigma}_L$ and $\widehat{\sigma}_U$ are standard errors of $\widehat{\theta}_U$ and $\widehat{\theta}_L$, respectively, and $\widehat{\Delta}$ is the estimated length of the identification region. Let

$$\text{CI}_\alpha^{\text{Theorem 2}} = \left[\widehat{\theta}_L - c_\alpha \widehat{\sigma}_L, \widehat{\theta}_U + c_\alpha \widehat{\sigma}_U \right], \quad (\text{F1})$$

where c_α solves

$$\Phi \left(c_\alpha + \frac{\widehat{\Delta}}{\max\{\widehat{\sigma}_L, \widehat{\sigma}_U\}} \right) - \Phi(-c_\alpha) = 1 - \alpha.$$

Since $\widehat{\theta}_L \leq \widehat{\theta}_U$ by construction, Lemma 3 and Proposition 1 of [Stoye \(2009\)](#) imply that $\theta_{\text{avg}} \in \text{CI}_\alpha^{\text{Theorem 2}}$ with probability $1 - \alpha$ uniformly as $n \rightarrow \infty$, provided that the data generating process satisfies mild regularity conditions given in Assumption 1 (i) and (ii) of [Stoye \(2009\)](#).

F.2. The case of Theorem 4. Recall that in this case, the sharp identified interval of θ_{avg} is given by $[\theta_L, \theta_{U_e}]$, where

$$\theta_{U_e} = \frac{\min\{1, \mathbb{P}(Y_i = 1 \mid Z_i = 1) + 1 - e(1)\} - \max\{0, \mathbb{P}(Y_i = 1 \mid Z_i = 0) - e(0)\}}{1 - \max\{0, \mathbb{P}(Y_i = 1 \mid Z_i = 0) - e(0)\}}.$$

It is convenient to introduce additional notation. Let

$$\xi_1 = \mathbb{P}(Y_i = 1 \mid Z_i = 1) + 1 - e(1) \quad \text{and} \quad \xi_2 = \mathbb{P}(Y_i = 1 \mid Z_i = 0) - e(0).$$

Then ξ_1 and ξ_2 can be estimated by their sample analogs as before: i.e.,

$$\widehat{\xi}_1 = \widehat{\mathbb{P}}(Y_i = 1 \mid Z_i = 1) + 1 - \widehat{e}(1) \quad \text{and} \quad \widehat{\xi}_2 = \widehat{\mathbb{P}}(Y_i = 1 \mid Z_i = 0) - \widehat{e}(0).$$

Furthermore, let $\widehat{\sigma}(\xi_1)$ and $\widehat{\sigma}(\xi_2)$ denote respective standard errors.²⁵ For any $\alpha \in (0, 1/2)$, define z_α such that $\Phi(z_\alpha) = 1 - \alpha$. Hence, $z_{\alpha/2}$ is the two-sided standard normal critical value. We now choose $\bar{\alpha}$ that is close to zero, say $\bar{\alpha} = 0.001$, by which we define the following three possibilities.

- (i) $\widehat{\xi}_1 - z_{\bar{\alpha}/4} \cdot \widehat{\sigma}(\xi_1) \geq 1$;
- (ii) $\widehat{\xi}_1 + z_{\bar{\alpha}/4} \cdot \widehat{\sigma}(\xi_1) \leq 1$ and $\widehat{\xi}_2 + z_{\bar{\alpha}/4} \cdot \widehat{\sigma}(\xi_2) \leq 0$;
- (iii) neither (i) nor (ii).

The critical value $z_{\bar{\alpha}/4}$ is used here to reflect the fact that the two equalities are tested jointly against both positive and negative directions, which can be viewed as pretesting. Case (i) corresponds to the case that $\widehat{\xi}_1$ is much larger than 1, implying that the upper bound is 1. Then we recommend using the confidence set such that $[\widehat{\theta}_L - z_{\alpha - \bar{\alpha}} \widehat{\sigma}_L, 1]$. To accommodate the error in the pretesting stage, we recommend using the $z_{\alpha - \bar{\alpha}}$ critical value here. Case (ii) suggests that $\widehat{\xi}_1$ is much smaller than 1 and $\widehat{\xi}_2$ is sufficiently less than zero. Therefore, in this subcase, the upper bound reduces to $\theta_{U_e} = \xi_1$. Then the estimators of both lower and upper bounds are asymptotically linear, implying that one can use the confidence interval similar to (F1). Again, to accommodate the error in the pretesting stage, we recommend using the $z_{\alpha - \bar{\alpha}}$ critical value in applying (F1). This ensures that the asymptotic coverage probability is at least as large as $1 - \alpha$ in applying (F1). It turns out that case (ii) was the relevant case for the empirical example reported Section 6.

Case (iii) is more complicated. We rely on a simple projection method to construct a valid confidence set.²⁶ It follows from the proofs of Lemma I6 and Theorem 4 that the two end points of the sharp identified interval of θ_{avg} have the form:

$$\max_{a,b \in [0,1]^2} \text{ and } \min_{a,b \in [0,1]^2} \frac{a-b}{1-b} \quad \text{s.t.} \quad \begin{cases} a \geq b, \\ a \in [\mathbb{P}(Y_i = 1 \mid Z_i = 1), \mathbb{P}(Y_i = 1 \mid Z_i = 1) + 1 - e(1)], \\ b \in [\mathbb{P}(Y_i = 1 \mid Z_i = 0) - e(0), \mathbb{P}(Y_i = 1 \mid Z_i = 0)]. \end{cases} \quad (\text{F2})$$

Rewrite the constraints above as

$$a \in [\mathbb{E}(\mathbb{1}\{Y_i = 1\} \mid Z_i = 1), \mathbb{E}(\mathbb{1}\{Y_i = 1\} \mid Z_i = 1) + \mathbb{E}(\mathbb{1}\{T_i = 0\} \mid Z_i = 1)] =: [m_a, M_a],$$

$$b \in [\mathbb{E}(\mathbb{1}\{Y_i = 1\} \mid Z_i = 0) - \mathbb{E}(\mathbb{1}\{T_i = 1\} \mid Z_i = 0), \mathbb{E}(\mathbb{1}\{Y_i = 1\} \mid Z_i = 0)] =: [m_b, M_b].$$

²⁵The standard errors can be constructed differently, depending on the sampling assumptions regarding the two marginal distributions of (Y_i, Z_i) and (T_i, Z_i) .

²⁶A similar inference method is adopted in Horowitz and Lee (2022); however, the subject matter is different here.

Since a and b are bounded by conditional expectations, it is simple to construct a joint confidence set for a and b in the form of the Cartesian product (e.g., using a Bonferroni correction). That is, we can find the confidence set such that the following holds asymptotically with probability at least $1 - \alpha$:

$$a \in [\hat{m}_a, \hat{M}_a] \quad \text{and} \quad b \in [\hat{m}_b, \hat{M}_b].$$

Consequently, the following optima include the true identified set of θ_{avg} defined in Equation (F2), asymptotically with probability at least $1 - \alpha$:

$$\max_{a,b \in [0,1]^2} \quad \text{and} \quad \min_{a,b \in [0,1]^2} \quad \frac{a-b}{1-b} \quad \text{s.t.} \quad \begin{cases} a \geq b, \\ a \in [\hat{m}_a, \hat{M}_a], \\ b \in [\hat{m}_b, \hat{M}_b]. \end{cases}$$

F.3. The case of Theorem 5. In this case, the sharp bound of θ_{avg} is given by $[\theta_L, 1]$. Thus, the one-sided confidence interval for θ_L provides the valid confidence set for θ_{avg} . That is, $\theta_{\text{avg}} \in [\hat{\theta}_L - z_\alpha \hat{\sigma}_L, 1]$ holds asymptotically with probability at least $1 - \alpha$, where z_α satisfies $\Phi(z_\alpha) = 1 - \alpha$.

APPENDIX G. INFERENCE ON THE LOCAL PERSUASION RATE

G.1. Case under Assumption C. Recall that in this case, θ_{local} is point identified by $\theta_{\text{local}} = \theta^*$, where

$$\theta^* := \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}{\mathbb{P}(Y_i = 0, T_i = 0 \mid Z_i = 0) - \mathbb{P}(Y_i = 0, T_i = 0 \mid Z_i = 1)}.$$

Thus, in this case, one can use the standard delta method to construct the two-sided confidence interval based on the asymptotic normality of the sample analog estimator of θ^* . Let $\hat{\theta}^*$ and $\hat{\sigma}^*$ denote the sample analog estimator of θ^* and its standard error, respectively. Then the random interval $[\hat{\theta}^* - z_{\alpha/2} \hat{\sigma}^*, \hat{\theta}^* + z_{\alpha/2} \hat{\sigma}^*]$ includes θ_{local} asymptotically with probability $1 - \alpha$.

G.2. Case under Assumption D. The sharp identified interval of θ_{local} is given by $[\theta_L^*, 1]$, where $\theta_L^* = \max(\text{Wald}, \theta_L)$ with

$$\text{Wald} := \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}{e(1) - e(0)}.$$

Thus, the lower bound on θ_{local} has the form of the intersection bound (Chernozhukov, Lee, and Rosen, 2013). To describe how to conduct inference, define the sample analog estimator of $\widehat{\text{Wald}}$:

$$\widehat{\text{Wald}} := \frac{\hat{\mathbb{P}}(Y_i = 1 \mid Z_i = 1) - \hat{\mathbb{P}}(Y_i = 1 \mid Z_i = 0)}{\hat{e}(1) - \hat{e}(0)}.$$

Let $\hat{\sigma}_{\text{Wald}}$ be its standard error. Then, the random interval $[\hat{\theta}_L^*(\alpha), 1]$ includes θ_{local} with probability at least $1 - \alpha$, where

$$\hat{\theta}_L^*(\alpha) := \max\{0, \widehat{\text{Wald}} - z_{\alpha/2} \hat{\sigma}_{\text{Wald}}, \hat{\theta}_L - z_{\alpha/2} \hat{\sigma}_L\}.$$

Here, the critical value $z_{\alpha/2}$ is based on simple Bonferroni correction. An adaptive inequality selection proposed in [Chernozhukov, Lee, and Rosen \(2013\)](#) can be adopted to construct a sharper critical value than $z_{\alpha/2}$.

G.3. Case under Assumption E. In this case, the sharp identified interval of θ_{local} coincides with that of θ_{avg} . Hence, one can use the inference method presented in [Appendix F.3](#).

APPENDIX H. SEMIPARAMETRIC ESTIMATION

H.1. Efficient Estimation. In this section, we are explicit about the vector X_i of exogenous covariates and consider semiparametrically efficient estimation of the two key parameters, i.e., θ_L and θ^* .²⁷ We focus on independent and identically distributed (i.i.d.) data again. For θ_L we work with the dataset $\{(Y_i, Z_i, X_i^\top)^\top : i = 1, 2, \dots, n\}$, whereas we use $\{(Y_i, T_i, Z_i, X_i^\top)^\top : i = 1, 2, \dots, n\}$ for θ^* . Since we are now explicit about X_i , the objects of interest will be defined by integrating X_i out as we explained in [appendix D.1](#): when we analyze the local persuasion rate, we will use the conditional distribution of X_i given the group of compliers. Let $\tilde{\theta}_L(x)$ be the sharp lower bound conditional on $X_i = x$ and let $\tilde{\theta}^*(x)$ be the local persuasion rate conditional on $X_i = x$: i.e., $\tilde{\theta}_L(x) = \tilde{\theta}_N(x)/\{1 - \tilde{\theta}_{L,D}(x)\}$ and $\tilde{\theta}^*(x) = \tilde{\theta}_N(x)/\tilde{\theta}_D^*(x)$, where

$$\tilde{\theta}_N(x) := \mathbb{P}(Y_i = 1 \mid Z_i = 1, X_i = x) - \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x),$$

$$\tilde{\theta}_{L,D}(x) := \mathbb{P}(Y_i = 1 \mid Z_i = 0, X_i = x),$$

$$\tilde{\theta}_D^*(x) := \mathbb{P}(Y_i = 0, T_i = 0 \mid Z_i = 0, X_i = x) - \mathbb{P}(Y_i = 0, T_i = 0 \mid Z_i = 1, X_i = x).$$

Now, we take care of X_i as we explained in [appendix D.1](#) to obtain the parameters we consider estimating in this section: i.e.,

$$\theta_L = \frac{\int \tilde{\theta}_N(x) dF(x)}{1 - \int \tilde{\theta}_{L,D}(x) dF(x)}, \quad \theta^* = \frac{\int \tilde{\theta}_N(x) dF(x)}{\int \tilde{\theta}_D^*(x) dF(x)}, \quad (\text{H1})$$

provided that $\int \tilde{\theta}_D^*(x) dF(x) \neq 0$ and $\int \tilde{\theta}_{L,D}(x) dF(x) < 1$.

Semiparametric estimation of θ_L and θ^* is straightforward: i.e.,

$$\hat{\theta}_L := \frac{n^{-1} \sum_{i=1}^n \hat{\theta}_N(X_i)}{1 - n^{-1} \sum_{i=1}^n \hat{\theta}_{L,D}(X_i)} \quad \text{and} \quad \hat{\theta}^* := \frac{\sum_{i=1}^n \hat{\theta}_N(X_i)}{\sum_{i=1}^n \hat{\theta}_D^*(X_i)}, \quad (\text{H2})$$

where $\hat{\theta}_N(x)$, $\hat{\theta}_{L,D}(x)$, $\hat{\theta}_D^*(x)$ are defined by replacing the probabilities in the definition of $\tilde{\theta}_N(x)$, $\tilde{\theta}_{L,D}(x)$, and $\tilde{\theta}_D^*(x)$ with their consistent estimators, respectively.

Semiparametric estimators like the ones in [Equation \(H2\)](#) converge at the usual \sqrt{n} rate. Instead of listing all regularity conditions, which are well understood in the literature (see, e.g., [Newey, 1994](#); [Ai](#)

²⁷Efficient estimation of θ_L^* or θ_{L_u} is significantly more challenging because of the min/max function. It also requires a careful construction of an estimator for the marginal persuasion rate. These are beyond the scope of the current paper but interesting topics for future research.

and Chen, 2003; Chen, Linton, and Van Keilegom, 2003; Ai and Chen, 2012; Akerberg, Chen, Hahn, and Liao, 2014), we will derive the pathwise derivatives of θ_L and θ^* . The theorem stated below is the first main result of this section.

Theorem H1. *Suppose that $e(0, x) < e(1, x)$ with $\inf_x e(0, x) > 0$ and $\sup_x e(1, x) < 1$, and that $0 < \inf_x \mathbb{P}(Z_i = 1 \mid X_i = x) \leq \sup_x \mathbb{P}(Z_i = 1 \mid X_i = x) < 1$. Then, the numerators and denominators of θ_L and θ^* are all pathwise differentiable in the sense of Newey (1994).*

Theorem H1 does not display the specific forms of the pathwise derivatives simply because their expressions are too long and distracting; they are provided in Lemmas I11 and I14.²⁸ Below we discuss the relevance of Theorem H1.

The pathwise differentiability can tell us a couple of things about the semiparametric estimators $\hat{\theta}_L$ and $\hat{\theta}^*$. Specifically, we can combine the pathwise derivatives of the numerators and denominators of θ_L and θ^* with the delta method, and we can find out what the influence functions $F_L(Y, Z, X)$ and $F^*(Y, T, Z, X)$ of $\hat{\theta}_L$ and $\hat{\theta}^*$ should be if they are asymptotically linear: see Lemmas I12 and I15 for the explicit expressions of $F_L(Y, T, Z, X)$ and $F^*(Y, T, Z, X)$. Therefore, the asymptotic variance of $\hat{\theta}_L$ and $\hat{\theta}^*$ will be $\mathbb{V}\{F_L(Y_i, Z_i, X_i)\}$ and $\mathbb{V}\{F^*(Y_i, T_i, Z_i, X_i)\}$, respectively. Further, we show that (any linear combination of) the pathwise derivatives of the numerator and denominator of θ_L is contained in the appropriate tangent space and that the same is true for θ^* . What this means is that $\mathbb{V}\{F_L(Y_i, Z_i, X_i)\}$ and $\mathbb{V}\{F^*(Y_i, T_i, Z_i, X_i)\}$ are in fact the semiparametric efficiency bounds of θ_L and θ^* , respectively. We summarize these implications in the following theorem.

Theorem H2. *Suppose that the conditions stated in Theorem H1 are satisfied. If $\hat{\theta}_L$ and $\hat{\theta}^*$ are \sqrt{n} -consistent and asymptotically linear, then their asymptotic variances are given by*

$$V_L = \mathbb{E}\{F_L^2(Y_i, Z_i, X_i)\} \quad \text{and} \quad V^* = \mathbb{E}\{F^{*2}(Y_i, T_i, Z_i, X_i)\}.$$

Further, V_L and V^* are the semiparametric efficiency bounds for estimating θ_L and θ^* , respectively.

By using the formulas of F_L and F^* provided in Lemmas I12 and I15, we can consistently estimate V_L and V^* . Alternatively, one can simply rely on some resampling techniques such as the bootstrap. Once we obtain the estimates \hat{V}_L and \hat{V}^* of the asymptotic variances, we can conduct inference on θ_L and θ^* . In fact, inference on θ_L can be naturally extended to that of θ_{avg} . Consider the simplest case where there is no uncertainty in the upper bound: i.e., we only have data on $(Y_i, Z_i, X_i^\top)^\top$. In this case, the sharp identified bounds of θ_{avg} is simply $[\theta_L, 1]$, the asymptotically valid confidence interval for θ_{avg} with the shortest length will be $[\hat{\theta}_L - z_\alpha \times \sqrt{\hat{V}_L/n}, 1]$, where z_α is the one-sided critical value from the standard normal distribution (e.g., when $\alpha = 0.05$, we have $z_\alpha = 1.645$). When we observe all of $(Y_i, T_i, Z_i, X_i^\top)^\top$,

²⁸The regularity conditions imposed on the exposure rates in Theorem H1 may not be necessary but we do not explore minimal conditions here.

θ^* is point identified and therefore the shortest confidence interval of θ^* can be obtained by using the usual two-sided critical value as in $\hat{\theta}^* \pm z_{\alpha/2} \times \sqrt{\hat{V}^*/n}$.

H.2. Revisiting the NTV Example. We revisit the NTV example in Appendix D.2 to illustrate the results of Appendix H.1. For simplicity, we focus on estimating the lower bound θ_L using a two-step parametric approach. For this exercise, we first create a binary instrument

$$Z_i := \mathbb{1}\{\text{Signal Power}_i > \text{median}(\text{Signal Power})\}$$

by using the original continuous instrument. The conditional lower bound $\tilde{\theta}_L(x)$ is estimated using probit models that are linear in covariates used in Appendix D.2 and is averaged out with respect to covariates by sample survey weight. The standard error is obtained by replacing population quantities of $F_L(Y_i, Z_i, X_i)$ in Lemma I12 with parametric estimates.

TABLE H1. Persuasion Rates: NTV Effects Using a Binary Instrument

	(1)	(2)
	Not voting for Unity	Voting for OVR
Point estimate of the lower bound	0.203	0.056
Standard error of the lower bound	0.072	0.024
One-sided 95% confidence interval for θ_{avg}	[0.084,1]	[0.017,1]

The estimation results are summarized in Table H1. Both lower bounds are significantly different from zero; however, they are far smaller than the estimates of θ_{avg} based on the original continuous instrument. This again illustrates the limitation of the identifying power of a binary instrument.

PART IV. PROOFS

APPENDIX I. PROOFS

I.1. Fréchet–Hoeffding Inequalities.

Lemma I1. For any events A, B and for any probability measure \mathbb{P}^* , we have

$$\max\{0, \mathbb{P}^*(A) - \mathbb{P}^*(B^c)\} \leq \mathbb{P}^*(A \cap B) \leq \min\{\mathbb{P}^*(A), \mathbb{P}^*(B)\},$$

where the bounds are sharp in that $\mathbb{P}^*(A \cap B)$ can be anything between the bounds without changing $\mathbb{P}^*(A)$ and $\mathbb{P}^*(B)$.

Proof. This is a version of the Fréchet–Hoeffding bounds. We prove it here for self-containment. The upper bound is trivially true. For the lower bound, simply note that

$$\mathbb{P}^*(A) \leq \mathbb{P}^*(A \cap B) + \mathbb{P}^*(B^c) \quad \text{and} \quad \mathbb{P}^*(B) \leq \mathbb{P}^*(A \cap B) + \mathbb{P}^*(A^c),$$

where $\mathbb{P}^*(A) - \mathbb{P}^*(B^c) = \mathbb{P}^*(B) - \mathbb{P}^*(A^c)$. For sharpness, note that the upper bound is achieved when $A \subset B$ or $B \subset A$. Also, the lower bound is achieved when $A \cap B = \emptyset$ or $A \cup B = \Omega$, where Ω is the entire sample space. To show that anything between the bounds can be achieved, consider the canonical probability space without loss of generality: i.e., $\Omega = [0, 1]$ and \mathbb{P}^* be the Lebesgue measure on the Borel σ -algebra on Ω . Choose $p_A, p_B \in [0, 1]$, where we assume that $p_A \geq p_B$ without loss of generality; the other case is symmetric. Let $A = [0, p_A]$ and $B = x + [p_A - p_B, p_A]$, where $0 \leq x \leq 1 - p_A$. So, A, B are events in Ω with $\mathbb{P}^*(A) = p_A, \mathbb{P}^*(B) = p_B$ for all $x \in [0, 1 - p_A]$. Now, note that $\mathbb{P}^*(A \cap B) = \max(p_B - x, 0)$ is a continuous function in $x \in [0, 1 - p_A]$, where its maximum and the minimum are given by p_B and $\max(p_B - 1 + p_A, 0)$. \square

I.2. Proof of Lemma 1. First of all, applying Lemma 11 with $P^*(A) = \mathbb{P}\{Y(1) = 1\}$ and $P^*(B) = \mathbb{P}\{Y(0) = 0\}$ yields Equation (3). For Equation (4), we show that $\mathbb{P}\{Y_i(1) = 1\} - \mathbb{P}\{Y_i(0) = 1\} = \mathbb{P}\{Y_i(1) = 1, Y_i(0) = 0\}$ if and only if $Y_i(1) \geq Y_i(0)$ with probability one. First, suppose that $Y_i(1) \geq Y_i(0)$. Then, $Y_i(1) - Y_i(0) = \mathbb{1}\{Y_i(1) = 1, Y_i(0) = 0\}$ with probability one, and hence taking the expectation on both sides shows the claim. Conversely, if $\mathbb{P}\{Y_i(1) = 1\} - \mathbb{P}\{Y_i(0) = 1\} = \mathbb{P}\{Y_i(1) = 1, Y_i(0) = 0\}$, then the fact that $\mathbb{P}\{Y_i(1) = 1\} = \mathbb{P}\{Y_i(1) = 1, Y_i(0) = 1\} + \mathbb{P}\{Y_i(1) = 1, Y_i(0) = 0\}$ and $\mathbb{P}\{Y_i(0) = 1\} = \mathbb{P}\{Y_i(1) = 1, Y_i(0) = 1\} + \mathbb{P}\{Y_i(1) = 0, Y_i(0) = 1\}$ shows that $\mathbb{P}\{Y_i(1) = 0, Y_i(0) = 1\} = 0$. \square

I.3. Proof of Lemma A1. If $q_i(0) \leq q_i(1)$, then $Y_i(0) = 1$ and $Y_i(1) = 0$ cannot happen. Now, conversely, suppose that $Y_i(0) \leq Y_i(1)$. If $\mathbb{P}\{q_i(1) < q_i(0)\} > 0$, then Assumption A1 implies that $\mathbb{P}\{q_i(1) < 1/(1 + U_i) < q_i(0)\} > 0$. This contradicts $\mathbb{P}\{Y_i(0) \leq Y_i(1)\} = 1$. \square

I.4. Proof of Theorem 1. By Assumption B,

$$\mathbb{P}\{Y_i(z) = 1\} = \mathbb{P}(Y_i(z) = 1 \mid Z_i = z) = \mathbb{P}(Y_i = 1 \mid T_i = z) = \mathbb{P}(Y_i = 1 \mid Z_i = z).$$

So, the assertion follows from Lemma 1 and the definition of θ_L . \square

I.5. Using Assumptions A to C.

Lemma I2. Suppose that Assumptions B and C hold. For $z = 0, 1$, $\mathbb{P}\{Y_i(1) = 1, V_i \leq e(z)\}$ is identified by

$$\mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = z).$$

Similarly, $\mathbb{P}\{Y_i(0) = 1, V_i > e(z)\}$ is identified by

$$\mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = z).$$

Proof. It follows from Assumption B. \square

Lemma I3. Suppose that Assumptions **B** and **C** hold. Then, $\mathbb{P}\{Y_i(1) = 1 \mid e(0) < V_i \leq e(1)\}$ is identified by

$$\frac{\mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 0)}{e(1) - e(0)}.$$

Similarly, $\mathbb{P}\{Y_i(0) = 1 \mid e(0) < V_i \leq e(1)\}$ is identified by

$$\frac{\mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) - \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 1)}{e(1) - e(0)}.$$

Proof. The first assertion follows from Lemma **I2** because

$$\mathbb{P}\{Y_i(1) = 1, e(0) < V_i \leq e(1)\} = \mathbb{P}\{Y_i(1) = 1, V_i \leq e(1)\} - \mathbb{P}\{Y_i(1) = 1, V_i \leq e(0)\}.$$

The second statement is similar. □

Lemma I4. Suppose that Assumptions **B** and **C** hold. Then,

$$\begin{aligned} \mathbb{P}\{Y_i(1) = 1\} &= \mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) + \mathbb{P}\{Y_i(1) = 1, V_i > e(1)\}, \\ \mathbb{P}\{Y_i(0) = 1\} &= \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) + \mathbb{P}\{Y_i(0) = 1, V_i \leq e(0)\}. \end{aligned}$$

Proof. For the first assertion, note that

$$\begin{aligned} \mathbb{P}\{Y_i(1) = 1\} &= \mathbb{P}\{Y_i(1) = 1 \mid e(0) < V_i \leq e(1)\}\{e(1) - e(0)\} \\ &\quad + \mathbb{P}\{Y_i(1) = 1, V_i \leq e(0)\} + \mathbb{P}\{Y_i(1) = 1, V_i > e(1)\}. \end{aligned} \quad (\text{I1})$$

By Lemmas **I2** and **I3**, the first two terms on the right-hand side of Equation **(I1)** are identified, and their sum is equal to $\mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1)$. The second assertion is similar. □

Lemma I5. Suppose that Assumptions **A** to **C** hold. Then, the sharp identified interval of $\mathbb{P}\{Y_i(1) = 1\}$ is given by

$$[\mathbb{P}(Y_i = 1 \mid Z_i = 1), \mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) + 1 - e(1)].$$

Similarly, the sharp identified interval of $\mathbb{P}\{Y_i(0) = 1\}$ is given by

$$[\mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0), \mathbb{P}(Y_i = 1 \mid Z_i = 0)].$$

Proof. For the first assertion, recall from lemma **I4** that

$$\mathbb{P}\{Y_i(1) = 1\} = \mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) + \mathbb{P}\{Y_i(1) = 1, V_i > e(1)\}, \quad (\text{I2})$$

where

$$\mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 1) = \mathbb{P}\{Y_i(0) = 1, V_i > e(1)\} \leq \mathbb{P}\{Y_i(1) = 1, V_i > e(1)\} \leq 1 - e(1).$$

Here, the first inequality is sharp under Assumption A and so is the second one by Lemma I1. Therefore, the sharp bounds of the second term on the right-hand side of Equation (I2) is the interval between $\mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 1)$ and $1 - e(1)$. Combining all these proves the first assertion. The second assertion is similar. \square

I.6. **Proof of Theorem 2.** Let $a = \mathbb{P}\{Y_i(1) = 1\}$ and $b = \mathbb{P}\{Y_i(0) = 1\}$: so, $\theta_{\text{avg}} = (a - b)/(1 - b)$. Let m_a, M_a be the lower and upper bounds of a provided in Lemma I5. Similarly, let m_b, M_b be the bounds of b given in Lemma I5. By Lemma I5 and the fact that the dependence between $Y_i(0)$ and $Y_i(1)$ is unrestricted except that $Y_i(0) \leq Y_i(1)$, the identified upper bound of θ_{avg} can be obtained by

$$\max_{a,b} \frac{a-b}{1-b} \quad \text{subject to} \quad a \in [m_a, M_a], b \in [m_b, M_b], a \geq b.$$

We can find the identified lower bound by solving minimization instead of maximization. Here, note that the restriction $a \geq b$ is redundant, because

$$\begin{aligned} m_a - M_b &= \mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0) \\ &= \mathbb{P}\{Y_i(1) = 1, e(0) < V_i \leq e(1)\} - \mathbb{P}\{Y_i(0) = 1, e(0) < V_i \leq e(1)\} \geq 0, \end{aligned}$$

where the last inequality is from $Y_i(1) \geq Y_i(0)$. So, the minimum is $\theta_L = (m_a - M_b)/(1 - M_b) \geq 0$ and the maximum is $\theta_U = (M_a - m_b)/(1 - m_b) \leq 1$: both θ_L and θ_U are well-defined because $m_b \leq M_b < 1$. Finally, sharpness follows from the intermediate value theorem because $(a - b)/(1 - b)$ varies continuously between θ_L and θ_U . \square

I.7. **Proof of Theorem 3.** Lemma I4 shows that

$$\mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) \leq \mathbb{P}\{Y_i(1) = 1\} \leq \mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) + 1 - e(1) \leq 1, \quad (\text{I3})$$

$$\mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) \leq \mathbb{P}\{Y_i(0) = 1\} \leq \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) + e(0) \leq 1 \quad (\text{I4})$$

if no restrictions are imposed on $\mathbb{P}\{Y_i(1) = 1, V_i > e(1)\}$ and $\mathbb{P}\{Y_i(0) = 1, V_i \leq e(0)\}$: these bounds are sharp. Therefore, combining (I3), (I4) with (3) in Lemma 1 and following the same reasoning as the proof of Theorem 2 shows the first assertion. We remark that the denominator in the upper bound is not equal to zero because $\mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) \leq \mathbb{P}(Y_i = 1 \mid Z_i = 0) < 1$. Also, if $\mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) + e(0) \uparrow 1$, then the lower bound goes to $\max(0, -\infty) = 0$ because $\mathbb{P}(Y_i = 1 \mid Z_i = 1) < 1$. Therefore, the lower bound is well-defined as well. For the second assertion, suppose that the stated stochastic dominance condition holds in addition. Then, it follows from Lemma I4 that

$$\mathbb{P}(Y_i = 1 \mid Z_i = 1) \leq \mathbb{P}\{Y_i(1) = 1\} \leq \mathbb{P}(Y_i = 1, T_i = 1 \mid Z_i = 1) + 1 - e(1) \leq 1, \quad (\text{I5})$$

$$\mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) \leq \mathbb{P}\{Y_i(0) = 1\} \leq \mathbb{P}(Y_i = 1, T_i = 0 \mid Z_i = 0) + e(0) \leq 1 \quad (\text{I6})$$

where the bounds are sharp. Using (I5), (I6), and (3) in Lemma 1, we follow the same reasoning as the proof of Theorem 2 to obtain the second result of the theorem. Again, the bounds are all well-defined because $\mathbb{P}(Y_i = 1 \mid Z_i = 0) < 1$. \square

I.8. Using Assumptions A, B and D.

Lemma I6. *Suppose that Assumptions A, B and D hold. The sharp identified interval of $\mathbb{P}\{Y_i(1) = 1\}$ is given by*

$$[\mathbb{P}(Y_i = 1 \mid Z_i = 1), \min\{1, \mathbb{P}(Y_i = 1 \mid Z_i = 1) + 1 - e(1)\}]$$

Similarly, the sharp identified interval of $\mathbb{P}\{Y_i(0) = 1\}$ is given by

$$[\max\{0, \mathbb{P}(Y_i = 1 \mid Z_i = 0) - e(0)\}, \mathbb{P}(Y_i = 1 \mid Z_i = 0)].$$

Proof. We start from Lemma I5, and we apply Lemma I1 to address the difference between Assumptions C and D. \square

I.9. **Proof of Theorem 4.** Similarly to the proof of Theorem 2, we need to consider

$$\max_{a,b} \text{ and } \min_{a,b} \frac{a-b}{1-b} \quad \text{subject to} \quad a \in [m_a, \tilde{M}_a], \quad b \in [\tilde{m}_b, M_b], \quad a \geq b,$$

where $m_a, \tilde{M}_a, \tilde{m}_b, M_b$ are given in Lemma I6. Follow the same reasoning as Theorem 2. \square

I.10. **Proof of Theorem 5.** Since Theorem 4 uses more information but its lower bound only depends on the distribution of (Y_i, Z_i) , it suffices to focus on the upper bound. From Theorem 4, we can find the sharp upper bound in this case by

$$\max_{0 \leq e(0) \leq e(1) \leq 1} \theta_{U_e} = \frac{\min\{1, \mathbb{P}(Y_i = 1 \mid Z_i = 1) + 1 - e(1)\} - \max\{0, \mathbb{P}(Y_i = 1 \mid Z_i = 0) - e(0)\}}{1 - \max\{0, \mathbb{P}(Y_i = 1 \mid Z_i = 0) - e(0)\}}.$$

Note that setting $e(0) = \mathbb{P}(Y_i = 1 \mid Z_i = 0) \leq \mathbb{P}(Y_i = 1 \mid Z_i = 1) = e(1)$ yields the maximum value 1. Sharpness follows from the fact that θ_{U_e} is continuous in $(e(0), e(1))$. \square

I.11. For the Compliers.

Proof of Theorem 6: For part (i), note that

$$\mathbb{P}(Y_i = 1 \mid Z_i = z) = \mathbb{P}\{Y_i(1) = 1, V_i \leq e(z)\} + \mathbb{P}\{Y_i(0) = 1, V_i > e(z)\},$$

from which it follows that

$$\theta_{\text{local}} = \frac{\mathbb{P}(Y_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i = 1 \mid Z_i = 0)}{\{e(1) - e(0)\} \mathbb{P}\{Y_i(0) = 0 \mid e(0) < V_i \leq e(1)\}}. \quad (\text{I7})$$

Finally, note that the denominator on the right-hand side of Equation (I7) is equal to

$$\mathbb{P}\{Y_i(0) = 0, e(0) < V_i \leq e(1)\} = \mathbb{P}\{Y_i(0) = 0, e(0) < V_i\} - \mathbb{P}\{Y_i(0) = 0, e(1) < V_i\},$$

where $\mathbb{P}\{Y_i(0) = 0, e(z) < V_i\} = \mathbb{P}\{Y_i = 0, T_i = 0 \mid Z_i = z\}$.

For part (ii), we look for sharp bounds for $\mathbb{P}\{Y_i(0) = 0, e(0) < V_i \leq e(1)\}$ under Assumption D. Using the fact that the sharp bounds of $\mathbb{P}(A \cap B \cap C)$ when $\mathbb{P}(A \cap B)$, $\mathbb{P}(B \cap C)$, and $\mathbb{P}(C \cap A)$ are given are equal to the interval between 0 and $\min\{\mathbb{P}(A \cap B), \mathbb{P}(B \cap C), \mathbb{P}(C \cap A)\}$, we know that

$$\begin{aligned} 0 &\leq \mathbb{P}\{Y_i(0) = 0, e(0) < V_i \leq e(1)\} \\ &\leq \min[\mathbb{P}\{Y_i(0) = 0, V_i > e(0)\}, e(1) - e(0), \mathbb{P}\{Y_i(0) = 0, V_i \leq e(1)\}], \end{aligned} \quad (18)$$

where it suffices to look for the sharp upper bound of the expression on the utmost right-hand side. First,

$$\mathbb{P}\{Y_i(0) = 0, V_i > e(0)\} = \mathbb{P}(Y_i = 0, T_i = 0 \mid Z_i = 0) \leq \mathbb{P}(Y_i = 0 \mid Z_i = 0),$$

where the inequality holds with equality when $\mathbb{P}(Y_i = 0, T_i = 1 \mid Z_i = 0) = 0$. Second, note that

$$\mathbb{P}\{Y_i(0) = 0, V_i \leq e(1)\} = \mathbb{P}\{Y_i(0) = 0, T_i = 1 \mid Z_i = 1\}$$

is totally unidentified. So, we conclude that the sharp upper bound of the term on the right-hand side of Equation (18) is

$$\min\{\mathbb{P}(Y_i = 0 \mid Z_i = 0), e(1) - e(0)\}.$$

The bound in part (iii) corresponds to the case where $e(1) - e(0) = 1$. □

Proof of Theorem 7: By the same reasoning as Lemma 1, we have

$$\theta_{\text{marginal}}(v) = \frac{\mathbb{E}\{Y_i(1) - Y_i(0) \mid V_i = v\}}{\mathbb{P}\{Y_i(0) = 0 \mid V_i = v\}}.$$

Then as shown in Heckman and Vytlacil (2005),

$$\mathbb{E}\{Y_i(1) - Y_i(0) \mid V_i = v\} = \left. \frac{\partial \mathbb{P}\{Y_i = 1 \mid e(Z_i) = e\}}{\partial e} \right|_{e=v}.$$

Also, by the same argument as in Heckman and Vytlacil (2005),

$$\mathbb{E}\{Y_i(0) \mid V_i = v\} = - \left. \frac{\partial \mathbb{P}\{Y_i = 1, T_i = 0 \mid e(Z_i) = e\}}{\partial e} \right|_{e=v}.$$

The desired result follows immediately. □

I.12. Proofs with Nonbinary Outcomes.

Proof of Lemma B1: Fixing $Y_{i0}(t) = 0$ for $t = 0, 1$, there are only four possibilities, where one of them can be ruled out by $Y_{i1}(1) \geq Y_{i1}(0)$: this is illustrated in the following table, where the outcomes in the third row (those with *) have probability zero.

Therefore,

$$\mathbb{P}\{Y_i(1) = (0, 1, 0), Y_i(0) = (0, 0, 1)\} = \mathbb{P}\{Y_i(1) = (0, 1, 0)\} - \mathbb{P}\{Y_i(0) = (0, 1, 0)\},$$

$Y(1)$	$Y(0)$
$(0, 1, 0)$	$(0, 1, 0)$
$(0, 1, 0)$	$(0, 0, 1)$
$(0, 0, 1)^*$	$(0, 1, 0)^*$
$(0, 0, 1)$	$(0, 0, 1)$

which takes care of the numerator of the conditional probability of θ_{mult} . For the denominator, note that

$$\mathbb{P}\{Y_i(0) = (0, 0, 1)\} = \mathbb{P}\{Y_{i,-1}(0) = 1\} = 1 - \mathbb{P}\{Y_{i0}(0) = 1\} - \mathbb{P}\{Y_{i1}(0) = 1\}. \quad \square$$

Lemma I7. For $j \in \{0, 1, -1\}$, we have

$$\mathbb{P}\{Y_{ij}(1) = 1 \mid e(0) < V_i \leq e(1)\} = \frac{p_j(1, 1 \mid 1) - p_j(1, 1 \mid 0)}{e(1) - e(0)}.$$

Similarly,

$$\mathbb{P}\{Y_{ij}(0) = 1 \mid e(0) < V_i \leq e(1)\} = \frac{p_j(1, 0 \mid 0) - p_j(1, 0 \mid 1)}{e(1) - e(0)}.$$

Proof. It follows from the same proof as Lemma I3. □

Lemma I8. For $z = 0, 1$ and $j \in \{0, 1, -1\}$, we have

$$\mathbb{P}\{Y_{ij}(1) = 1, V_i \leq e(z)\} = p_j(1, 1 \mid z).$$

Similarly,

$$\mathbb{P}\{Y_{ij}(0) = 1, V_i > e(z)\} = p_j(1, 0 \mid z).$$

Proof. It follows from the same proof as Lemma I2. □

Lemma I9. The sharp identified region of $(\mathbb{P}\{Y_{i1}(1) = 1\}, \mathbb{P}\{Y_{i0}(0) = 1\}, \mathbb{P}\{Y_{i1}(0) = 1\})$ is given by $\mathcal{S}_1(1) \times \mathcal{S}_{01}(0)$, where $\mathcal{S}_1(1) = [p_1(1 \mid 1), p_1(1, 1 \mid 1) + 1 - e(1)]$ is the sharp identified interval of $\mathbb{P}\{Y_{i1}(1) = 1\}$ and

$$\begin{aligned} \mathcal{S}_{01}(0) = \left\{ (x, y) \in [0, 1]^2 : p_0(1, 0 \mid 0) \leq x \leq p_0(1, 0 \mid 0) + e(0), \right. \\ \left. p_1(1, 0 \mid 0) \leq y \leq p_1(1 \mid 0), \quad x + y \leq \sum_{j=0}^1 p_j(1, 0 \mid 0) + e(0) \right\} \neq \emptyset \end{aligned}$$

is the sharp identified region of $(\mathbb{P}\{Y_{i0}(0) = 1\}, \mathbb{P}\{Y_{i1}(0) = 1\})$.

Proof. Note that

$$\begin{aligned} \mathbb{P}\{Y_{i1}(1) = 1\} = \mathbb{P}\{Y_{i1}(1) = 1, V_i \leq e(0)\} + \mathbb{P}\{Y_{i1}(1) = 1, e(0) < V_i \leq e(1)\} \\ + \mathbb{P}\{Y_{i1}(1) = 1, V_i > e(1)\}, \end{aligned}$$

where the first two terms on the right-hand side are identified by Lemmas 17 and 18. Specifically,

$$\mathbb{P}\{Y_{i1}(1) = 1\} = p_1(1, 1 | 1) + \mathbb{P}\{Y_{i1}(1) = 1, V_i > e(1)\}. \quad (\text{I9})$$

Similarly, for $j \in \{0, 1, -1\}$, we obtain

$$\mathbb{P}\{Y_{ij}(0) = 1\} = p_j(1, 0 | 0) + \mathbb{P}\{Y_{ij}(0) = 1, V_i \leq e(0)\}. \quad (\text{I10})$$

Here, note that there is no restriction on the relationship between $\tilde{a} := \mathbb{P}\{Y_{i1}(1) = 1, V_i > e(1)\}$ and $(\tilde{b}, \tilde{c}) := (\mathbb{P}\{Y_{i0}(0) = 1, V_i \leq e(0)\}, \mathbb{P}\{Y_{i1}(0) = 1, V_i \leq e(0)\})$. Therefore, it suffices to consider \tilde{a} and (\tilde{b}, \tilde{c}) separately, after which we take the Cartesian product of the two sharp identified sets. For the sharp interval of \tilde{a} , we simply combine Equation (I9) with

$$\mathbb{P}(Y_{i1} = 1, T_i = 0 | Z_i = 1) = \mathbb{P}\{Y_{i1}(0) = 1, V_i > e(1)\} \leq \mathbb{P}\{Y_{i1}(1) = 1, V_i > e(1)\} \leq 1 - e(1), \quad (\text{I11})$$

which yields $\mathbb{P}\{Y_{i1}(1) = 1\} \in \mathcal{S}_1(1)$. For the sharp region of (\tilde{b}, \tilde{c}) , we use Equation (I10) with the fact that

$$0 \leq \mathbb{P}\{Y_{i1}(0) = 1, V_i \leq e(0)\} \leq \mathbb{P}\{Y_{i1}(1) = 1, V_i \leq e(0)\} = \mathbb{P}(Y_{i1} = 1, T_i = 1 | Z_i = 0).$$

Therefore, it follows that

$$\mathbb{P}\{Y_{i0}(0) = 1\} \in [p_0(1, 0 | 0), p_0(1, 0 | 0) + e(0)], \quad (\text{I12})$$

$$\mathbb{P}\{Y_{i1}(0) = 1\} \in [p_1(1, 0 | 0), p_1(1 | 0)], \quad (\text{I13})$$

$$\mathbb{P}\{Y_{i,-1}(0) = 1\} \in [p_{-1}(1, 0 | 0), p_{-1}(1, 0 | 0) + e(0)], \quad (\text{I14})$$

where we must have

$$\sum_{j \in \{0, 1, -1\}} \mathbb{P}\{Y_{ij}(0) = 1\} = 1 \quad \text{and} \quad \sum_{j \in \{0, 1, -1\}} p_j(1, 0 | 0) = 1 - e(0). \quad (\text{I15})$$

by Assumption B1.²⁹ Therefore, we can rewrite Equation (I14) by using Equation (I15). Specifically, Equation (I14) can be written as

$$\begin{aligned} \mathbb{P}\{Y_{i,-1}(0) = 1\} &= 1 - \mathbb{P}\{Y_{i0}(0) = 1\} - \mathbb{P}\{Y_{i1}(0) = 1\} \\ &\in [1 - e(0) - p_0(1, 0 | 0) - p_1(1, 0 | 0), 1 - e(0) - p_0(1, 0 | 0) - p_1(1, 0 | 0) + e(0)], \end{aligned}$$

²⁹The upper end point of the bounds in Equation (I13) is no larger than the lower end point of the interval $\mathcal{S}_1(1)$ because

$$p_1(1 | 1) - p_1(1 | 0) = \mathbb{P}\{Y_{i1}(1) = 1, e(0) < V_i \leq e(1)\} - \mathbb{P}\{Y_{i1}(0) = 1, e(0) < V_i \leq e(1)\} \geq 0,$$

where the inequality is by $Y_{i1}(1) \geq Y_{i1}(0)$.

which implies

$$\mathbb{P}\{Y_{i0}(0) = 1\} + \mathbb{P}\{Y_{i1}(0) = 1\} \in \left[\sum_{j=0}^1 p_j(1,0 | 0), \sum_{j=0}^1 p_j(1,0 | 0) + e(0) \right]. \quad (\text{I16})$$

But the lower bound in Equation (I16) is implied by Equations (I12) and (I13), whereas the upper bound is not redundant. Note that $\mathcal{S}_{01}(0)$ is not empty, unless

$$-p_0(1,0 | 0) + \sum_{j=0}^1 p_j(1,0 | 0) + e(0) < p_1(1,0 | 0),$$

which is not possible. \square

Proof of Theorem B2: Let $a := \mathbb{P}\{Y_{i1}(1) = 1\}$, $b := \mathbb{P}\{Y_{i0}(0) = 1\}$, and $c := \mathbb{P}\{Y_{i1}(0) = 1\}$, and we have

$$\theta_{\text{mult}} = \frac{a - c}{1 - b - c},$$

where the sharp identified region of (a, b, c) is $\mathcal{S}_1(1) \times \mathcal{S}_{01}(0)$ by Lemma I9. Therefore, we can obtain the sharp identified bounds of θ_{mult} by solving constrained maximization/minimization problems:

$$\max / \min_{a,b,c} \frac{a - c}{1 - b - c} \quad \text{s.t.} \quad \begin{cases} m_a \leq a \leq M_a, \\ m_b \leq b \leq M_b, \\ m_c \leq c \leq M_c, \\ b + c \leq M_{b+c}, \end{cases} \quad (\text{I17})$$

where m_a, M_a, \dots, M_{b+c} are given in the definitions of $\mathcal{S}_1(1)$ and $\mathcal{S}_{01}(0)$. Here, the restriction of $a \geq c$ is automatic, because $m_a = p_1(1 | 1) \geq p_1(1 | 0) = M_c$ as we explained in Footnote 29. Further, note that $m_b + m_c \leq M_{b+c}$ because $e(0) \geq 0$; therefore, the set of (b, c) that satisfies the constraints is not empty. In the following arguments, the set of the feasible values of (b, c) is important, which is illustrated in Figure I1.

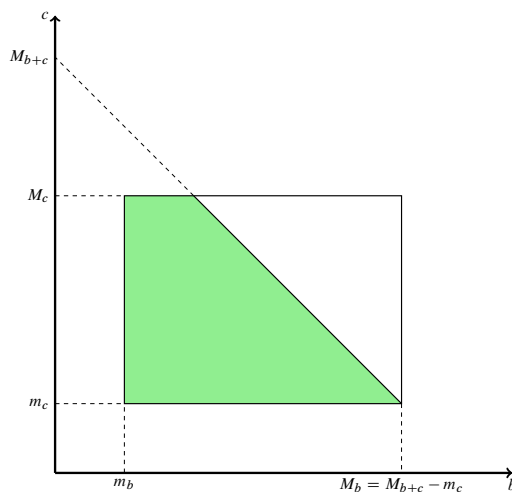
Consider the minimization first, for which it suffices to solve

$$\min_{b,c} \frac{m_a - c}{1 - b - c} \quad \text{s.t.} \quad \begin{cases} m_b \leq b \leq M_b, \\ m_c \leq c \leq M_c, \\ b + c \leq M_{b+c}. \end{cases}$$

Here, the objective function is monotonically increasing in b for every c . Therefore,

$$\min \left(\frac{m_a - c}{1 - m_b - c}, 1 \right) \leq \min \left(\frac{m_a - c}{1 - b - c}, 1 \right) \leq \theta_{\text{mult}} \leq 1 \quad (\text{I18})$$

for all $m_c \leq c \leq \min(M_{b+c} - m_b, M_c) = M_c$, where the utmost left-hand side inequality in Equation (I18) holds with equality when $b = m_b$.

FIGURE I1. The set of the feasible values of (b, c) 

Note: $M_{b+c} - m_b = p_0(1,0 | 0) + p_1(1,0 | 0) + e(0) - p_0(1,0 | 0) \geq p_1(1,0 | 0) + p_1(1,1 | 0) = M_c$.

Note: $M_{b+c} - m_c = p_0(1,0 | 0) + e(0) = M_b$.

Note: $M_{b+c} - M_b = p_0(1,0 | 0) + p_1(1,0 | 0) + e(0) - p_0(1,0 | 0) - e(0) \leq p_1(1,0 | 0) + p_1(1,1 | 0) = M_c$.

Now, we consider two possibilities: i.e., $m_a + m_b \geq 1$ and $m_a + m_b < 1$. If $m_a + m_b \geq 1$, then the utmost left-hand side of Equation (I18) becomes 1 for any $c \leq 1 - m_b$. But, taking $c = m_c$ leads to $m_c \leq M_{b+c} - m_b \leq 1 - m_b$ by

$$M_{b+c} = p_0(1,0 | 0) + p_1(1,0 | 0) + e(0) = 1 - e(0) - p_{-1}(1,0 | 0) + e(0) \leq 1.$$

Therefore, we conclude $\theta_{\text{mult}} = 1$ in this case, which is achieved when $a = m_a, b = m_b$, and $c = m_c$.

If $m_a + m_b < 1$, then $(m_a - c)/(1 - m_b - c)$ is decreasing in c , and therefore, we simply take the largest value of c to achieve the minimum: i.e.,

$$\frac{m_a - M_c}{1 - m_b - M_c} \leq \theta_{\text{mult}}.$$

This is indeed the sharp lower bound, because $(a - c)/(1 - b - c)$ is a continuous function in (a, b, c) and the lower bound is achieved when $a = m_a, b = m_b$, and $c = M_c$. In particular, Assumption B1 ensures that $m_a - M_c \geq 0$, and since we assume that $p_{-1}(1 | 0) > 0$ together with Assumption B1, we know that $0 \leq m_b + M_c = p_0(1,0 | 0) + p_1(1 | 0) \leq p_0(1 | 0) + p_1(1 | 0) = 1 - p_{-1}(1 | 0) < 1$.

The sharp upper bound can be similarly obtained by solving the maximization problem in Equation (I17), for which it suffices to consider

$$\max_{b,c} \frac{M_a - c}{1 - b - c} \quad \text{s.t.} \quad \begin{cases} m_b \leq b \leq M_b, \\ m_c \leq c \leq M_c, \\ b + c \leq M_{b+c}. \end{cases}$$

Since the objective function is increasing in b , it suffices to have b take its largest value within the feasible set, i.e., $b = \min(M_{b+c} - m_c, M_b) = M_b$.

Then, we have

$$\theta_{\text{mult}} \leq \frac{M_a - c}{1 - b - c} \leq \frac{M_a - c}{1 - M_b - c}, \quad (\text{I19})$$

for all $m_c \leq c \leq \min(M_{b+c} - m_b, M_c) = M_c$, where the utmost right-hand side inequality in Equation (I19) holds with equality when $b = M_b$.

Now, we consider two possibilities: i.e., $M_a + M_b \geq 1$ and $M_a + M_b < 1$. First, suppose that $M_a + M_b \geq 1$. Then, the utmost right-hand side of Equation (I19) is no smaller than 1 for any $c \leq 1 - M_b$. Therefore, the sharp upper bound of θ_{mult} is trivial and equal to 1 in this case. Suppose that $M_a + M_b < 1$. Then, the utmost right-hand side expression in Equation (I19) is decreasing in c , and therefore

$$\theta_{\text{mult}} \leq \frac{M_a - c}{1 - b - c} \leq \frac{M_a - m_c}{1 - M_b - m_c}.$$

Sharpness follows from the fact that $(a - c)/(1 - b - c)$ is continuous in (a, b, c) and the upper bound is achieved when $a = M_a, b = M_b$, and $c = m_c$. We note that this bound is smaller than 1, because we have $M_a - m_c < 1 - M_b - m_c$ in this case, where $M_a - m_c = p_1(1, 1 | 1) + 1 - e(1) - p_1(1, 0 | 0) \geq 1 - e(1) + \mathbb{P}\{Y_{i1}(0) = 1, e(0) < V_i \leq e(1)\} > 0$.

Finally, since $m_a + m_b \leq M_a + M_b$, what we have shown so far yields the sharp upper and lower bounds in all of the three possible cases, i.e., (i) $1 \leq m_a + m_b$, (ii) $m_a + m_b < 1 \leq M_a + M_b$, and (iii) $m_a + m_b \leq M_a + M_b < 1$. \square

Proof of Theorem B3: We use Lemmas I1 and I9 to obtain the sharp identified region of $(\mathbb{P}\{Y_{i1}(1) = 1\}, \mathbb{P}\{Y_{i0}(0) = 1\}, \mathbb{P}\{Y_{i1}(0) = 1\})$, after which we follow the same reasoning as in the proof of Theorem B2. \square

Proof of Theorem B4: The case of $p_1(1 | 1) = 1$ is trivial, because we can simply take $e(0) = p_0(1 | 0)$; this leads us to case (i) in Theorem B3, but the claimed bounds yield the same conclusion. Suppose that $p_1(1 | 1) < 1$. Then, we can always choose $e(1) = \max\{p_0(1 | 0), p_1(1 | 0)\}$ and take $e(0)$ smaller than but arbitrarily close to $e(1)$, which leads us to case (ii) in Theorem B3 with the claimed bounds. \square

I.13. Proofs for the Results for Efficient Estimation of θ_L . In this part of the appendix we consider the efficiency issues for the θ_L parameter defined in Equation (H1). So, we assume that the data available to us are i.i.d. observations of $(Y_i, Z_i, X_i^\top)^\top$. Below we write f for F' , i.e., the density of X_i . Further, we use the following notation:

$$\begin{aligned} \mathcal{P}_{y_1|z_1}(x) &:= \mathbb{P}(Y_i = 1 | Z_i = 1, X_i = x) & \text{and} & & \mathcal{P}_{y_0|z_1}(x) &:= 1 - \mathcal{P}_{y_1|z_1}(x), \\ \mathcal{P}_{y_1|z_0}(x) &:= \mathbb{P}(Y_i = 1 | Z_i = 0, X_i = x) & \text{and} & & \mathcal{P}_{y_0|z_0}(x) &:= 1 - \mathcal{P}_{y_1|z_0}(x), \\ \mathcal{P}_{z_1}(x) &:= \mathbb{P}(Z_i = 1 | X_i = x) & \text{and} & & \mathcal{P}_{z_0}(x) &:= 1 - \mathcal{P}_{z_1}(x). \end{aligned}$$

Then, we can write the likelihood function as follows.

$$q_L(y, z, x) = f(x) \prod_{j \in \{1,0\}} \left[\mathcal{P}_{z_j}(x) \{ \mathcal{P}_{y_1|z_j}(x)^y \mathcal{P}_{y_0|z_j}(x)^{1-y} \}^{\tilde{z}_j} \right],$$

where $\tilde{z}_1 = z$ and $\tilde{z}_0 = 1 - z$. Therefore, the loglikelihood function is given by

$$\log q_L(y, z, x) = \log f(x) + \sum_{j \in \{1,0\}} \{ \tilde{z}_j \log \mathcal{P}_{z_j}(x) + \tilde{z}_j y \log \mathcal{P}_{y_1|z_j}(x) + \tilde{z}_j (1 - y) \log \mathcal{P}_{y_0|z_j}(x) \}. \quad (\text{I20})$$

Lemma I10. *The tangent space for θ_L is given by*

$$\mathcal{T}_L = \left\{ \alpha_f(x) + \{z - \mathcal{P}_{z_1}(x)\} \alpha_z(x) + \sum_{j \in \{0,1\}} \tilde{z}_j \{y - \mathcal{P}_{y_1|z_j}(x)\} \alpha_{y|z_j}(x) \right\},$$

where α_f is any square-integrable function with $\mathbb{E}\{\alpha_f(X_i)\} = 0$ and $\alpha_z, \alpha_{y|z_j}$ are any square-integrable functions of x .

Proof. Let $\mathcal{P}_{z_j}(x | \gamma), \mathcal{P}_{y_k|z_j}(x | \gamma)$ denote regular parametric submodels indexed by γ :³⁰ we will denote the true value by γ_0 . Then, it follows from Equation (I20) that the score is given by

$$s(y, z, x | \gamma_0) = s_X(x | \gamma_0) + s_{Z|X}(z, x | \gamma_0) + s_{Y|Z,X}(y, z, x | \gamma_0), \quad (\text{I21})$$

where

$$\begin{aligned} s_X(x | \gamma_0) &= \frac{1}{f(x)} \frac{\partial f(x | \gamma)}{\partial \gamma}, \\ s_{Z|X}(z, x | \gamma_0) &= \left\{ \frac{z}{\mathcal{P}_{z_1}(x)} - \frac{1-z}{\mathcal{P}_{z_0}(x)} \right\} \frac{\partial \mathcal{P}_{z_1}(x | \gamma)}{\partial \gamma}, \\ s_{Y|Z,X}(y, z, x | \gamma_0) &= \sum_{j \in \{1,0\}} \tilde{z}_j \left\{ \frac{y}{\mathcal{P}_{y_1|z_j}(x)} - \frac{1-y}{\mathcal{P}_{y_0|z_j}(x)} \right\} \frac{\partial \mathcal{P}_{y_1|z_j}(x | \gamma)}{\partial \gamma}, \end{aligned}$$

where all the derivatives are evaluated at γ_0 . The conclusion follows from the fact that all the derivatives are unrestricted here. \square

For $j = 0, 1$, let $\beta_j(\gamma) := \int \mathcal{P}_{y_1|z_j}(x | \gamma) f(x | \gamma) dx$, and let $\beta_{j0} = \beta_j(\gamma_0)$.

Lemma I11. *For any $(c_0, c_1) \neq (0, 0)$, $c_0 \beta_0(\gamma) + c_1 \beta_1(\gamma)$ is pathwise differentiable, and its pathwise derivative is given by $c_0 F_0(Y, Z, X) + c_1 F_1(Y, Z, X)$, where for $j = 0, 1$,*

$$F_j(Y, Z, X) := \left(\frac{Z \{Y - \mathcal{P}_{y_1|z_1}(X)\}}{\mathcal{P}_{z_1}(X)} \right)^j \left(\frac{(1-Z) \{Y - \mathcal{P}_{y_1|z_0}(X)\}}{\mathcal{P}_{z_0}(X)} \right)^{1-j} + \mathcal{P}_{y_1|z_j}(X) - \beta_{j0}$$

Proof. First,

$$\frac{\partial \beta_j(\gamma_0)}{\partial \gamma} = \int \frac{\partial \mathcal{P}_{y_1|z_j}(x | \gamma_0)}{\partial \gamma} f(x) dx + \int \mathcal{P}_{y_1|z_j}(x) \frac{\partial f(x | \gamma_0)}{\partial \gamma} dx.$$

³⁰There is no loss of generality in assuming that γ is a scalar.

Now, note that $s(Y, Z, X \mid \gamma_0)$ given in Equation (I21) is the sum of Bernoulli scores, and it follows that

$$\frac{\partial \beta_j(\gamma_0)}{\partial \gamma} = \mathbb{E}\{F_j(Y, Z, X)s(Y, Z, X \mid \gamma_0)\}. \quad \square$$

Lemma I12. *The efficient influence function of θ_L is given by*

$$F_L(Y, Z, X) := \frac{1}{1 - \beta_{00}} \left\{ F_1(Y, Z, X) - \frac{1 - \beta_{10}}{1 - \beta_{00}} F_0(Y, Z, X) \right\}.$$

Proof. For any $(c_1, c_0) \neq (0, 0)$, $c_1 F_1 + c_0 F_0$ is in the tangent space \mathcal{T}_L : i.e., its projections onto \mathcal{T}_L is just itself. Also, the scores given in Equation (I21) can approximate any mean zero random variable with an arbitrarily small mean squared error. Therefore, Theorem 2.1 in Newey (1994) and lemma I11 ensure that the efficient influence function of $\sum_{j=0}^1 c_j \beta_j$ is given by $\sum_{j=0}^1 c_j F_j(Y, Z, X)$. So, the efficient influence function of $[\beta_{10} \ \beta_{00}]^\top$ is given by $[F_1(Y, Z, X) \ F_0(Y, Z, X)]^\top$. We now note that $\theta_L = (\beta_{10} - \beta_{00}) / \{1 - \beta_{00}\}$, and we apply the delta method. \square

I.14. Proofs for the Results for Efficient Estimation of θ^* . We now derive the efficient influence function of the local persuasion parameter defined in Equation (H1) when an i.i.d. sample of (Y_i, T_i, Z_i, X_i) is available. Similarly to the previous subsection we use the following notation:

$$\mathcal{P}_{z_1}(x) := \mathbb{P}(Z_i = 1 \mid X_i = x) \quad \text{and} \quad \mathcal{P}_{z_0} := 1 - \mathcal{P}_{z_1}(x),$$

$$\mathcal{P}_{t_1|z_1}(x) := \mathbb{P}(T_i = 1 \mid Z_i = 1, X_i = x) \quad \text{and} \quad \mathcal{P}_{t_0|z_1}(x) := 1 - \mathcal{P}_{t_1|z_1}(x),$$

$$\mathcal{P}_{t_1|z_0}(x) := \mathbb{P}(T_i = 1 \mid Z_i = 0, X_i = x) \quad \text{and} \quad \mathcal{P}_{t_0|z_0}(x) := 1 - \mathcal{P}_{t_1|z_0}(x),$$

$$\mathcal{P}_{y_1|t_1,z_1}(x) := \mathbb{P}(Y_i = 1 \mid T_i = 1, Z_i = 1, X_i = x) \quad \text{and} \quad \mathcal{P}_{y_0|t_1,z_1}(x) := 1 - \mathcal{P}_{y_1|t_1,z_1}(x),$$

$$\mathcal{P}_{y_1|t_0,z_1}(x) := \mathbb{P}(Y_i = 1 \mid T_i = 0, Z_i = 1, X_i = x) \quad \text{and} \quad \mathcal{P}_{y_0|t_0,z_1}(x) := 1 - \mathcal{P}_{y_1|t_0,z_1}(x),$$

$$\mathcal{P}_{y_1|t_1,z_0}(x) := \mathbb{P}(Y_i = 1 \mid T_i = 1, Z_i = 0, X_i = x) \quad \text{and} \quad \mathcal{P}_{y_0|t_1,z_0}(x) := 1 - \mathcal{P}_{y_1|t_1,z_0}(x),$$

$$\mathcal{P}_{y_1|t_0,z_0}(x) := \mathbb{P}(Y_i = 1 \mid T_i = 0, Z_i = 0, X_i = x) \quad \text{and} \quad \mathcal{P}_{y_0|t_0,z_0}(x) := 1 - \mathcal{P}_{y_1|t_0,z_0}(x).$$

Using this notation, the likelihood function can be written as

$$\begin{aligned} q^*(y, t, z, x) &= f(x) \prod_{j=\{1,0\}} \left[\mathcal{P}_{z_j}(x) \prod_{k=\{1,0\}} \{ \mathcal{P}_{t_k|z_j}(x) \prod_{\ell=\{1,0\}} \mathcal{P}_{y_\ell|t_k,z_j}(x)^{\tilde{y}_\ell} \}^{\tilde{t}_k} \right]^{\tilde{z}_j} \\ &= f(x) \prod_{j=\{1,0\}} \left[\mathcal{P}_{z_j}(x)^{\tilde{z}_j} \prod_{k=\{1,0\}} \{ \mathcal{P}_{t_k|z_j}(x)^{\tilde{t}_k \tilde{z}_j} \prod_{\ell=\{1,0\}} \mathcal{P}_{y_\ell|t_k,z_j}(x)^{\tilde{y}_\ell \tilde{t}_k \tilde{z}_j} \} \right], \end{aligned}$$

where $\tilde{z}_1 = z, \tilde{z}_0 = 1 - z, \tilde{t}_1 = t, \tilde{t}_0 = 1 - t$, and $\tilde{y}_1 = y, \tilde{y}_0 = 1 - y$. Therefore, the loglikelihood function is given by

$$\begin{aligned} \log q^*(y, t, z, x) &= \log f(x) + \sum_{j \in \{1,0\}} \tilde{z}_j \log \mathcal{P}_{z_j}(x) \\ &\quad + \sum_{j \in \{1,0\}} \sum_{k \in \{0,1\}} \tilde{t}_k \tilde{z}_j \log \mathcal{P}_{t_k|z_j}(x) + \sum_{\ell \in \{1,0\}} \sum_{j \in \{1,0\}} \sum_{k \in \{0,1\}} \tilde{y}_\ell \tilde{t}_k \tilde{z}_j \log \mathcal{P}_{y_\ell|t_k z_j}(x). \end{aligned} \quad (\text{I22})$$

Lemma I13. *The tangent space for θ^* is given by*

$$\begin{aligned} \mathcal{T}^* &= \left\{ \alpha_f(x) + \{z - \mathcal{P}_{z_1}(x)\} \alpha_z(x) + \sum_{j \in \{0,1\}} \tilde{z}_j \{t - \mathcal{P}_{t_1|z_j}(x)\} \alpha_{t|z_j}(x) \right. \\ &\quad \left. + \sum_{j \in \{0,1\}} \sum_{k \in \{0,1\}} \tilde{t}_k \tilde{z}_j \{y - \mathcal{P}_{y_1|t_k z_j}(x)\} \alpha_{y|t_k z_j}(x) \right\}, \end{aligned}$$

where α_f is any square-integrable function with $\mathbb{E}\{\alpha_f(X_1)\} = 0$, and $\alpha_z, \alpha_{t|z_j}, \alpha_{y|t_k z_j}$ are any square-integrable functions of x .

Proof. Let $\mathcal{P}_{z_j}(x | \gamma), \mathcal{P}_{t_k|z_j}(x | \gamma), \mathcal{P}_{y_\ell|t_k z_j}(x | \gamma)$ denote regular parametric submodels indexed by γ : as in the proof of Lemma I10, γ is a scalar-valued parameter and its true value is denoted by γ_0 . From the loglikelihood function described in Equation (I22), we know that the score of the regular parametric submodel can be written as follows:

$$\begin{aligned} s^*(y, t, z, x | \gamma_0) &= \frac{1}{f(x)} \frac{\partial f(x | \gamma)}{\partial \gamma} + \left\{ \frac{\tilde{z}_1}{\mathcal{P}_{z_1}(x)} - \frac{\tilde{z}_0}{\mathcal{P}_{z_0}(x)} \right\} \frac{\partial \mathcal{P}_{z_1}(x | \gamma)}{\partial \gamma} \\ &\quad + \sum_{j \in \{1,0\}} \tilde{z}_j \left\{ \frac{\tilde{t}_1}{\mathcal{P}_{t_1|z_j}(x)} - \frac{\tilde{t}_0}{\mathcal{P}_{t_0|z_j}(x)} \right\} \frac{\partial \mathcal{P}_{t_1|z_j}(x | \gamma)}{\partial \gamma} \\ &\quad + \sum_{j \in \{1,0\}} \sum_{k \in \{1,0\}} \tilde{t}_k \tilde{z}_j \left\{ \frac{\tilde{y}_1}{\mathcal{P}_{y_1|t_k z_j}(x)} - \frac{\tilde{y}_0}{\mathcal{P}_{y_0|t_k z_j}(x)} \right\} \frac{\partial \mathcal{P}_{y_1|t_k z_j}(x | \gamma)}{\partial \gamma}, \end{aligned} \quad (\text{I23})$$

where all the derivatives are evaluated at γ_0 . Here, we can do further algebra by using

$$\begin{aligned} \frac{\partial \mathcal{P}_{z_1}(x | \gamma)}{\partial \gamma} &= -\frac{\partial \mathcal{P}_{z_0}(x | \gamma)}{\partial \gamma}, \\ \frac{\partial \mathcal{P}_{t_1|z_j}(x | \gamma)}{\partial \gamma} &= -\frac{\partial \mathcal{P}_{t_0|z_j}(x | \gamma)}{\partial \gamma} \quad \text{for } j = 0, 1 \\ \frac{\partial \mathcal{P}_{y_1|t_k z_j}(x | \gamma)}{\partial \gamma} &= -\frac{\partial \mathcal{P}_{y_0|t_k z_j}(x | \gamma)}{\partial \gamma} \quad \text{for } j, k = 0, 1. \end{aligned}$$

Therefore, the score in equation (I23) can be rewritten as follows:

$$s^*(y, t, z, x | \gamma_0) = s_X(x | \gamma_0) + s_{Z|X}(z, x | \gamma_0) + s_{T|Z,X}(t, z, x | \gamma_0) + s_{Y|T,Z,X}(y, t, z, x | \gamma_0), \quad (\text{I24})$$

where

$$s_X(x | \gamma_0) := \frac{1}{f(x)} \frac{\partial f(x | \gamma)}{\partial \gamma},$$

$$\begin{aligned}
 s_{Z|X}(z, x | \gamma_0) &:= \frac{z - \mathcal{P}_{z_1}(x)}{\mathcal{P}_{z_1}(x)\{1 - \mathcal{P}_{z_1}(x)\}} \frac{\partial \mathcal{P}_{z_1}(x | \gamma)}{\partial \gamma}, \\
 s_{T|Z,X}(t, z, x | \gamma_0) &:= \sum_{j \in \{1,0\}} \frac{\tilde{z}_j \{t - \mathcal{P}_{t_1|z_j}(x)\}}{\mathcal{P}_{t_1|z_j}(x)\{1 - \mathcal{P}_{t_1|z_j}(x)\}} \frac{\partial \mathcal{P}_{t_1|z_j}(x | \gamma)}{\partial \gamma}, \\
 s_{Y|T,Z,X}(y, t, z, x | \gamma_0) &:= \sum_{j \in \{1,0\}} \sum_{k \in \{1,0\}} \frac{\tilde{t}_k \tilde{z}_j \{y - \mathcal{P}_{y_1|t_k,z_j}(x)\}}{\mathcal{P}_{y_1|t_k,z_j}(x)\{1 - \mathcal{P}_{y_1|t_k,z_j}(x)\}} \frac{\partial \mathcal{P}_{y_1|t_k,z_j}(x | \gamma)}{\partial \gamma}.
 \end{aligned}$$

Interpretation of each term should be straightforward. For example, $s_{T|Z,X}(t, z, x | \gamma_0)$ is the score of T at t conditional on $Z = z, X = x$. Finally, the conclusion follows from Equation (I24) and the fact that all the derivatives here are unrestricted. \square

Now, define

$$\begin{aligned}
 \beta_N^* &:= \int Q_N(x) f(x) dx, \\
 \beta_D^* &:= \int Q_D(x) f(x) dx,
 \end{aligned}$$

where

$$\begin{aligned}
 Q_N(x) &:= \sum_{k \in \{1,0\}} \mathcal{P}_{y_1|t_k,z_1}(x) \mathcal{P}_{t_k|z_1}(x) - \sum_{k \in \{1,0\}} \mathcal{P}_{y_1|t_k,z_0}(x) \mathcal{P}_{t_k|z_0}(x), \\
 Q_D(x) &:= \mathcal{P}_{y_0|t_0,z_0}(x) \mathcal{P}_{t_0|z_0}(x) - \mathcal{P}_{y_0|t_0,z_1}(x) \mathcal{P}_{t_0|z_1}(x).
 \end{aligned}$$

Then, θ^* given in Equation (H1) is equal to β_N^* / β_D^* . Below we derive the efficient influence functions of β_N^* and β_D^* jointly, after which we will use the delta method to obtain that of θ^* . Let $\beta_N^*(\gamma), \beta_D^*(\gamma), Q_N(x | \gamma)$, and $Q_D(x | \gamma)$ be the perturbed versions of $\beta_N^*, \beta_D^*, Q_N(x)$, and $Q_D(x)$ along regular parametric sub-models, respectively.

Define the following functions:

$$\begin{aligned}
 F_{1N}^*(Y, T, Z, X) &= \frac{TZ}{\mathcal{P}_{t_1|z_1}(X) \mathcal{P}_{z_1}(X)} \{Y - \mathcal{P}_{y_1|t_1,z_1}(X)\} \mathcal{P}_{t_1|z_1}(X) \\
 F_{2N}^*(Y, T, Z, X) &= \frac{(1-T)Z}{\mathcal{P}_{t_0|z_1}(X) \mathcal{P}_{z_1}(X)} \{Y - \mathcal{P}_{y_1|t_0,z_1}(X)\} \mathcal{P}_{t_0|z_1}(X) \\
 F_{3N}^*(Y, T, Z, X) &= -\frac{T(1-Z)}{\mathcal{P}_{t_1|z_0}(X) \mathcal{P}_{z_0}(X)} \{Y - \mathcal{P}_{y_1|t_1,z_0}(X)\} \mathcal{P}_{t_1|z_0}(X) \\
 F_{4N}^*(Y, T, Z, X) &= -\frac{(1-T)(1-Z)}{\mathcal{P}_{t_0|z_0}(X) \mathcal{P}_{z_0}(X)} \{Y - \mathcal{P}_{y_1|t_0,z_0}(X)\} \mathcal{P}_{t_0|z_0}(X) \\
 F_{5N}^*(Y, T, Z, X) &= \frac{Z}{\mathcal{P}_{z_1}(X)} \{T - \mathcal{P}_{t_1|z_1}(X)\} \{\mathcal{P}_{y_1|t_1,z_1}(X) - \mathcal{P}_{y_1|t_0,z_1}(X)\} \\
 F_{6N}^*(Y, T, Z, X) &= -\frac{1-Z}{\mathcal{P}_{z_0}(X)} \{T - \mathcal{P}_{t_1|z_0}(X)\} \{\mathcal{P}_{y_1|t_1,z_0}(X) - \mathcal{P}_{y_1|t_0,z_0}(X)\} \\
 F_{7N}^*(Y, T, Z, X) &= Q_N(X) - \beta_N^*
 \end{aligned}$$

Also, define

$$\begin{aligned}
F_{1D}^*(Y, T, Z, X) &= \frac{(1-T)Z}{\mathcal{P}_{t_0|z_1}(X)\mathcal{P}_{z_1}(X)} \{Y - \mathcal{P}_{y_1|t_0,z_1}(X)\} \mathcal{P}_{t_0|z_1}(X) \\
F_{2D}^*(Y, T, Z, X) &= -\frac{(1-T)(1-Z)}{\mathcal{P}_{t_0|z_0}(X)\mathcal{P}_{z_0}(X)} \{Y - \mathcal{P}_{y_1|t_0,z_0}(X)\} \mathcal{P}_{t_0|z_0}(X) \\
F_{3D}^*(Y, T, Z, X) &= \frac{Z}{\mathcal{P}_{z_1}(X)} \{T - \mathcal{P}_{t_1|z_1}(X)\} \mathcal{P}_{y_0|t_0,z_1}(X) \\
F_{4D}^*(Y, T, Z, X) &= -\frac{1-Z}{\mathcal{P}_{z_0}(X)} \{T - \mathcal{P}_{t_1|z_0}(X)\} \mathcal{P}_{y_0|t_0,z_0}(X) \\
F_{5D}^*(Y, T, Z, X) &= Q_D(X) - \beta_D^*.
\end{aligned}$$

Lemma I14. For any $(c_N, c_D) \neq (0, 0)$, $c_N \beta_N^*(\gamma) + c_D \beta_D^*(\gamma)$ is pathwise differentiable, and its pathwise derivative is given by $c_N F_N^*(Y, T, Z, X) + c_D F_D^*(Y, T, Z, X)$, where $F_N^*(Y, T, Z, X) := \sum_{j=1}^7 F_{jN}^*(Y, T, Z, X)$ and $F_D^*(Y, T, Z, X) := \sum_{j=1}^5 F_{jD}^*(Y, T, Z, X)$.

Proof. We first calculate the derivatives of the Q functions with respect to γ (evaluated at γ_0). For this calculation, there are only seven relevant derivatives, i.e.,

$$\begin{aligned}
&\frac{\partial \mathcal{P}_{y_1|t_1,z_1}(x|\gamma)}{\partial \gamma}, \frac{\partial \mathcal{P}_{y_1|t_0,z_1}(x|\gamma)}{\partial \gamma}, \frac{\partial \mathcal{P}_{y_1|t_1,z_0}(x|\gamma)}{\partial \gamma}, \frac{\partial \mathcal{P}_{y_1|t_0,z_0}(x|\gamma)}{\partial \gamma}, \\
&\frac{\partial \mathcal{P}_{t_1|z_1}(x|\gamma)}{\partial \gamma}, \frac{\partial \mathcal{P}_{t_1|z_0}(x|\gamma)}{\partial \gamma}, \\
&\frac{\partial f(x|\gamma)}{\partial \gamma}.
\end{aligned}$$

Indeed,

$$\begin{aligned}
\frac{\partial Q_N(x|\gamma)}{\partial \gamma} &= \sum_{k \in \{1,0\}} \frac{\partial \mathcal{P}_{y_1|t_k,z_1}(x|\gamma)}{\partial \gamma} \mathcal{P}_{t_k|z_1}(x) + \left\{ \mathcal{P}_{y_1|t_1,z_1}(x) - \mathcal{P}_{y_1|t_0,z_1}(x) \right\} \frac{\partial \mathcal{P}_{t_1|z_1}(x|\gamma)}{\partial \gamma} \\
&\quad - \sum_{k \in \{1,0\}} \frac{\partial \mathcal{P}_{y_1|t_k,z_0}(x|\gamma)}{\partial \gamma} \mathcal{P}_{t_k|z_0}(x) - \left\{ \mathcal{P}_{y_1|t_1,z_0}(x) - \mathcal{P}_{y_1|t_0,z_0}(x) \right\} \frac{\partial \mathcal{P}_{t_1|z_0}(x|\gamma)}{\partial \gamma},
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial Q_D(x|\gamma)}{\partial \gamma} &= \frac{\partial \mathcal{P}_{y_1|t_0,z_1}(x|\gamma)}{\partial \gamma} \mathcal{P}_{t_0|z_1}(x) - \frac{\partial \mathcal{P}_{y_1|t_0,z_0}(x|\gamma)}{\partial \gamma} \mathcal{P}_{t_0|z_0}(x) \\
&\quad + \mathcal{P}_{y_0|t_0,z_1} \frac{\partial \mathcal{P}_{t_1|z_1}(x|\gamma)}{\partial \gamma} - \mathcal{P}_{y_0|t_0,z_0} \frac{\partial \mathcal{P}_{t_1|z_0}(x|\gamma)}{\partial \gamma}.
\end{aligned}$$

We now consider

$$\partial \beta_N^*(\gamma_0) / \partial \gamma = \int \frac{\partial Q_N(x|\gamma)}{\partial \gamma} f(x) dx + \int Q_N(x) \frac{\partial f(x)}{\partial \gamma} dx,$$

which is the sum of the following terms:

$$\begin{aligned}
& \int \mathcal{P}_{t_1|z_1}(x) f(x) \frac{\partial \mathcal{P}_{y_1|t_1,z_1}(x | \gamma)}{\partial \gamma} dx, \\
& \int \mathcal{P}_{t_0|z_1}(x) f(x) \frac{\partial \mathcal{P}_{y_1|t_0,z_1}(x | \gamma)}{\partial \gamma} dx, \\
& - \int \mathcal{P}_{t_1|z_0}(x) f(x) \frac{\partial \mathcal{P}_{y_1|t_1,z_0}(x | \gamma)}{\partial \gamma} dx, \\
& - \int \mathcal{P}_{t_0|z_0}(x) f(x) \frac{\partial \mathcal{P}_{y_1|t_0,z_0}(x | \gamma)}{\partial \gamma} dx, \\
& \int \{ \mathcal{P}_{y_1|t_1,z_1}(x) - \mathcal{P}_{y_1|t_0,z_1}(x) \} f(x) \frac{\partial \mathcal{P}_{t_1|z_1}(x | \gamma)}{\partial \gamma} dx, \\
& - \int \{ \mathcal{P}_{y_1|t_1,z_0}(x) - \mathcal{P}_{y_1|t_0,z_0}(x) \} f(x) \frac{\partial \mathcal{P}_{t_1|z_0}(x | \gamma)}{\partial \gamma} dx, \\
& \int Q_N(x) \frac{\partial f(x | \gamma_0)}{\partial \gamma} dx.
\end{aligned}$$

Therefore, $F_N^*(Y, T, Z, X) := \sum_{j=1}^N F_{jN}^*(Y, T, Z, X)$ satisfies

$$\partial \beta_N^*(\gamma_0) / \partial \gamma = \mathbb{E}\{F_N^*(Y, T, Z, X) s^*(Y, T, Z, X | \gamma_0)\},$$

where s^* is defined in Equation (I24): in fact, this equation can be seen immediately from the fact that s^* is the sum of Bernoulli scores.³¹

Now, we turn to

$$\partial \beta_D^*(\gamma_0) / \partial \gamma = \int \frac{\partial Q_D(x | \gamma)}{\partial \gamma} f(x) dx + \int Q_D(x) \frac{\partial f(x | \gamma_0)}{\partial \gamma} dx,$$

which is the sum of the following terms:

$$\begin{aligned}
& \int \mathcal{P}_{t_0|z_1}(x) f(x) \frac{\partial \mathcal{P}_{y_1|t_0,z_1}(x | \gamma_0)}{\partial \gamma} dx \\
& - \int \mathcal{P}_{t_0|z_0}(x) f(x) \frac{\partial \mathcal{P}_{y_1|t_0,z_0}(x | \gamma_0)}{\partial \gamma} dx \\
& \int \mathcal{P}_{y_0|t_0,z_1}(x) f(x) \frac{\partial \mathcal{P}_{t_1|z_1}(x | \gamma_0)}{\partial \gamma} dx
\end{aligned}$$

³¹To see this point, it is helpful to consider a simple binary example. For instance, consider a generic binary variable B such that $\mathbb{P}(B = 1 | X, \gamma) = p_B(X | \gamma)$. Suppose that the parameter β satisfies $\partial \beta(\gamma) / \partial \gamma = \int A(x) \partial p_B(X | \gamma) / \partial \gamma f(x) dx = \mathbb{E}\{A(X) \partial p_B(X | \gamma) / \partial \gamma\}$. Here, the score of B given X is

$$s_B(B | X) = \frac{B - p_B(X)}{p_B(X)\{1 - p_B(X)\}} \frac{\partial p_B(X | \gamma)}{\partial \gamma}.$$

Now, we are looking for the function $G^*(B, X)$ such that

$$\mathbb{E}\{G^*(B, X) s_B(B | X)\} = \mathbb{E}\left\{A(X) \frac{\partial p_B(X | \gamma)}{\partial \gamma}\right\},$$

where

$$\mathbb{E}\{G^*(B, X) s_B(B | X)\} = \mathbb{E}\left[\{G^*(1, X) - G^*(0, X)\} \frac{\partial p_B(X | \gamma)}{\partial \gamma}\right].$$

Therefore, we immediately see that $G^*(B, X) = \{B - p_B(X)\} A(X)$ does the job.

$$- \int \mathcal{P}_{y_0|t_0,z_0}(x) f(x) \frac{\partial \mathcal{P}_{t_1|z_0}(x | \gamma_0)}{\partial \gamma} dx$$

$$\int Q_D(x) \frac{\partial f(x | \gamma_0)}{\partial \gamma} dx.$$

Therefore, $F_D(Y, T, Z, X) := \sum_{j=1}^5 F_{jD}^*(Y, T, Z, X)$ satisfies

$$\frac{\partial \beta_D^*(\gamma_0)}{\partial \gamma} = \mathbb{E}\{F_D^*(Y, T, Z, X) s^*(Y, T, Z, X | \gamma_0)\},$$

where s^* is defined in Equation (I24). □

Lemma I15. *The efficient influence function of θ^* is given by*

$$F^*(Y, T, Z, X) := \frac{1}{\beta_D^*} \{F_N^*(Y, T, Z, X) - \theta^* F_D^*(Y, T, Z, X)\}$$

Proof. For any $(c_N, c_D) \neq 0$, $c_N F_N^* + c_D F_D^*$ is in the tangent space \mathcal{T}^* : i.e., its projection onto \mathcal{T}^* is just itself. Also, the scores given in Equation (I24) can approximate any mean zero random variable with an arbitrarily small mean squared error. Therefore, Theorem 2.1 in Newey (1994) and lemma I14 ensure that the efficient influence function of $c_N \beta_N^* + c_D \beta_D^*$ is given by $c_N F_N^*(Y, T, Z, X) + c_D F_D^*(Y, T, Z, X)$. So, the efficient influence function of $[\beta_N \ \beta_D]^\top$ is given by $[F_N(Y, T, Z, X) \ F_D(Y, T, Z, X)]^\top$. We now note that $\theta^* = \beta_N^* / \beta_D^*$, and we apply the delta method. □

Proof of Theorem H1 It follows from Lemmas I11 and I14. □

Proof of Theorem H2 It follows from Lemmas I12 and I15. □

REFERENCES

- ACKERBERG, D., X. CHEN, J. HAHN, AND Z. LIAO (2014): "Asymptotic efficiency of semiparametric two-step GMM," *Review of Economic Studies*, 81(3), 919–943.
- AI, C., AND X. CHEN (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.
- (2012): "The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions," *Journal of Econometrics*, 170(2), 442–457.
- BASSI, V., AND I. RASUL (2017): "Persuasion: A Case Study of Papal Influence on Fertility Preferences and Behavior," *American Economic Journal: Applied Economics*, 9(4), 250–302.
- BERGEMANN, D., AND S. MORRIS (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57(1), 44–95.
- CALVI, R., A. LEWBEL, AND D. TOMMASI (2022): "LATE With Missing or Mismeasured Treatment," *Journal of Business & Economic Statistics*, 40(4), 1701–1717.
- CARNEIRO, P., J. J. HECKMAN, AND E. VYTLACIL (2011): "Estimating marginal returns to education," *American Economic Review*, 101(6), 2754–2781.

- CHE, Y.-K., W. DESSEIN, AND N. KARTIK (2013): "Pandering to Persuade," *American Economic Review*, 103(1), 47–79.
- CHEN, X., O. LINTON, AND I. VAN KEILEGOM (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71(5), 1591–1608.
- CHEN, Y., AND D. Y. YANG (2019): "The Impact of Media Censorship: 1984 or Brave New World?," *American Economic Review*, 109(6), 2294–2332.
- CHERNOZHUKOV, V., S. LEE, AND A. ROSEN (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.
- CROSS, P. J., AND C. F. MANSKI (2002): "Regressions, Short and Long," *Econometrica*, 70(1), 357–368.
- DEATON, A. (2010): "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 48(2), 424–55.
- DELLAVIGNA, S., R. ENIKOLOPOV, V. MIRONOVA, M. PETROVA, AND E. ZHURAVSKAYA (2014): "Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia," *American Economic Journal: Applied Economics*, 6(3), 103–32.
- DELLAVIGNA, S., AND M. GENTZKOW (2010): "Persuasion: empirical evidence," *Annual Review of Economics*, 2(1), 643–669.
- DELLAVIGNA, S., AND E. KAPLAN (2007): "The Fox News effect: Media bias and voting," *Quarterly Journal of Economics*, 122(3), 1187–1234.
- DELLAVIGNA, S., J. A. LIST, AND U. MALMENDIER (2012): "Testing for Altruism and Social Pressure in Charitable Giving," *Quarterly Journal of Economics*, 127(1), 1–56.
- ENIKOLOPOV, R., M. PETROVA, AND E. ZHURAVSKAYA (2011): "Media and Political Persuasion: Evidence from Russia," *American Economic Review*, 101(7), 3253–85.
- GENTZKOW, M. (2006): "Television and Voter Turnout," *Quarterly Journal of Economics*, 121(3), 931–972.
- GENTZKOW, M., AND E. KAMENICA (2017): "Competition in Persuasion," *Review of Economic Studies*, 84(1), 300–322.
- GENTZKOW, M., J. M. SHAPIRO, AND M. SINKINSON (2011): "The effect of newspaper entry and exit on electoral politics," *American Economic Review*, 101(7), 2980–3018.
- GERBER, A. S., D. KARLAN, AND D. BERGAN (2009): "Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions," *American Economic Journal: Applied Economics*, 1(2), 35–52.
- GREEN, D. P., AND A. S. GERBER (2000): "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment," *American Political Science Review*, 94(3), 653–663.
- (2001): "Getting out the youth vote: Results from randomized field experiments," Natural Field Experiments 00260, The Field Experiments Website.

- GREEN, D. P., A. S. GERBER, AND D. W. NICKERSON (2003): "Getting Out the Vote in Local Elections: Results from Six Door-to-Door Canvassing Experiments," *The Journal of Politics*, 65(4), 1083–1096.
- HECKMAN, J. J. (2010): "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 48(2), 356–98.
- HECKMAN, J. J., AND R. PINTO (2018): "Unordered Monotonicity," *Econometrica*, 86(1), 1–35.
- HECKMAN, J. J., J. SMITH, AND N. CLEMENTS (1997): "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts," *Review of Economic Studies*, 64(4), 487–535.
- HECKMAN, J. J., S. URZUA, AND E. VYTLACIL (2006): "Understanding instrumental variables in models with essential heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.
- (2008): "Instrumental variables in models with multiple outcomes: The general unordered case," *Annales d'économie et de statistique*, 91/92, 151–174.
- HECKMAN, J. J., AND E. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.
- (2007a): "Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 6 Part B, pp. 4779–4874. Elsevier.
- (2007b): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 6 Part B, pp. 4875–5143. Elsevier.
- HOROWITZ, J. L., AND S. LEE (2022): "Inference in a Class of Optimization Problems: Confidence Regions and Finite Sample Bounds on Errors in Coverage Probabilities," *Journal of Business & Economic Statistics*, p. in press.
- IMBENS, G., AND C. F. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.
- IMBENS, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48(2), 399–423.
- IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62(2), 467–475.
- IMBENS, G. W., AND D. B. RUBIN (1997): "Estimating outcome distributions for compliers in instrumental variables models," *Review of Economic Studies*, 64(4), 555–574.
- IMBENS, G. W., AND J. M. WOOLDRIDGE (2009): "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, 47(1), 5–86.

- KAMENICA, E., AND M. GENTZKOW (2011): "Bayesian Persuasion," *American Economic Review*, 101(6), 2590–2615.
- KIRBY, W. T. (2008): "The Public Sermon: Paul's Cross and the culture of persuasion in England, 1534–1570," *Renaissance and Reformation/Renaissance et Réforme*, pp. 3–29.
- LANDRY, C. E., A. LANGE, J. A. LIST, M. K. PRICE, AND N. G. RUPP (2006): "Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment," *Quarterly Journal of Economics*, 121(2), 747–782.
- LEE, S., AND B. SALANIÉ (2018): "Identifying Effects of Multivalued Treatments," *Econometrica*, 86(6), 1939–1963.
- MANSKI, C. F. (1997): "Monotone treatment response," *Econometrica*, pp. 1311–1334.
- (2003): *Partial Identification of Probability Distributions*. Springer-Verlag, New York.
- (2007): *Identification for Prediction and Decision*. Harvard University Press, New York.
- (2018): "Credible Ecological Inference for Medical Decisions with Personalized Risk Assessment," *Quantitative Economics*, 9(2), 541–569.
- MANSKI, C. F., AND J. V. PEPPER (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(4), 997–1010.
- MARTIN, G. J., AND A. YURUKOGLU (2017): "Bias in Cable News: Persuasion and Polarization," *American Economic Review*, 107(9), 2565–99.
- NEWWEY, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica*, pp. 1349–1382.
- NGUIMKEU, P., A. DENTEH, AND R. TCHERNIS (2019): "On the estimation of treatment effects with endogenous misreporting," *Journal of Econometrics*, 208(2), 487–506.
- PRAT, A. (2018): "Media Power," *Journal of Political Economy*, 126(4), 1747–1783.
- RAPP, C. (2010): "Aristotle's Rhetoric," in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta. Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, spring 2010 edn., <https://plato.stanford.edu/archives/spr2010/entries/aristotle-rhetoric/>.
- STOYE, J. (2009): "More on Confidence Regions for Partially Identified Parameters," *Econometrica*, 77(4), 1299–1315.
- TAMER, E. (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195.
- URA, T. (2018): "Heterogeneous Treatment Effects With Mismeasured Endogenous Treatment," *Quantitative Economics*, 9(3), 1335–1370.
- VYTLACIL, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70(1), 331–341.