IDENTIFICATION OF TREATMENT EFFECTS WITH SELECTIVE PARTICIPATION IN A RANDOMIZED TRIAL

BRENDAN KLINE AND ELIE TAMER

Abstract. Randomized trials (RTs) are used to learn about treatment effects. This paper studies identification of average treatment response (ATR) and average treatment effect (ATE) from RT data under various assumptions. The focus is the problem of external validity of the RT. RT data need not point identify the ATR or ATE because of selective participation in the RT. The paper reports partial identification and point identification results for the ATR and ATE based on RT data under a variety of assumptions. The results include assumptions sufficient to point identify the ATR or ATE from RT data. Under weaker assumptions, the ATR or ATE are partially identified. Further, attention is given to identification of the sign of the ATE and identification of whether participation in the RT is selective. Finally, identification from RT data is compared to identification from observational data.

Keywords: experiment, identification, randomized trial, treatment effect

University of Texas at Austin
Harvard University
E-mail addresses: brendan.kline@austin.utexas.edu, elietamer@fas.harvard.edu.

Date: April 2018. Based on an earlier paper distributed in January 2011 as "Using observational vs randomized controlled trial data to learn about treatment effects." We thank Chuck Manski and seminar participants at the University of Illinois at Urbana-Champaign for useful comments. We also thank the co-editor and two anonymous referees for helpful comments and suggestions. Financial support from the NSF is gratefully acknowledged. Any errors are ours.

1. Introduction

This paper is concerned with identification of treatment effects in a population of individuals each of whom is characterized by a response function $y_i(d) : d \in \mathcal{D}$ where d is a treatment that belongs to a finite set of mutually exclusive treatments $\mathcal{D}$. Particular emphasis is placed on identification of the average treatment response (ATR) of treatment d, which is

$$ATR(d) \equiv E(y_i(d)),$$

and the average treatment effect (ATE) of treatment d' versus treatment d, which is

$$ATE(d', d) \equiv E(y_i(d')) - E(y_i(d)).$$

Observational data is subject to treatment selection bias (or treatment endogeneity) when the response function is related to the realized treatment. For example, there can be concern that treatment choice is correlated with unobservables that are also correlated with the outcome. One possible solution to this is a randomized trial (RT) or randomized controlled trial (RCT). RTs and RCTs have come to be commonly known as the "gold standard" for statistical evidence. RTs are also used increasingly often in economics. See for example Banerjee and Duflo (2009) for a review of the use of experiments in development economics. The defining characteristic of an RT is that treatments are randomly assigned to participants, eliminating treatment selection bias.

This paper studies identification of ATR and ATE with RT data under a variety of assumptions. The focus is on *selection into an RT*. This concerns the way that individuals are invited to participate in an RT, the way that individuals decide whether to participate in an RT, and the population of interest. Even if the RT has internal validity, selective participation threatens the generalizability of the results from an RT beyond the sample of participants in the study, commonly known as the problem of *external validity*.

This paper builds on the fact that participation in an RT is a decision in the same way that treatment selection is a decision in observational data. The decision to participate in an RT may reflect an even more selective decision than the decision of which treatment to select in

observational data because of ethical and legal requirements on experiments involving human subjects. Indeed, the Declaration of Helsinki (e.g., World Medical Association (2013)) states that in medical research "the physician or another appropriately qualified individual must then seek the potential subject's freely-given informed consent, preferably in writing" and that "physicians who combine medical research with medical care should involve their patients in research only to the extent that this is justified by its potential preventive, diagnostic or therapeutic value and if the physician has good reason to believe that participation in the research study will not adversely affect the health of the patients who serve as research subjects."

The ethics of medical RTs is not settled but one standard view is "the uncertainty principle whereby randomisation to treatment is acceptable when an individual doctor is genuinely unsure which treatment is best for a patient" and another standard view is that "clinical equipoise, reflecting collective professional uncertainty over treatment, is the soundest ethical criterion" (from the debate in Weijer et al. (2000); note that one of these quotes is from one side of the debate and the other quote is from the other side). Similarly, although there does not yet exist a large literature on the ethics of RTs in economics, related ethical concerns would seem to apply to RTs in economics. This suggests that participation in RTs is selective because it is related to the response function. More generally, the same optimizing behavior that leads to treatment selection suggests optimizing behavior that leads to selective participation in RTs.

This paper focuses on external validity, which concerns identification of the ATR and ATE defined on the population of interest. Under the maintained assumptions, the RT is "ideal" in the sense that there are no threats to internal validity. Equivalently, the RT is assumed to point identify the ATR and ATE for the subpopulation participating in the RT. This means that possible problems relating to conducting an experiment for any given subpopulation of participants (e.g., treatment compliance) are assumed away in order to focus on external validity. In other words, this paper is not concerned with the possible

problems in experiments relating to whether the ATR and ATE can be point identified for the subpopulation participating in the RT.

Per the fundamental problem of causal inference, the identification problem studied in this paper relates to missing data because counterfactual outcomes are missing data. However, the model of an RT used in this paper results in a unique identification problem. Specifically, the model makes it possible to distinguish different sources of missing data on counterfactual outcomes. Those include 1) missing data due to the decision of whether an individual is invited to participate in the RT, 2) missing data that is due to the decision of the invitee of whether to participate in the RT, and 3) missing data due to the standard issue that any given individual can be observed to receive at most one of the treatments. Therefore, the setup and hence identification results differ from those in the related literature.

More specifically, this paper models the RT as a three step procedure, similar to the stylized model presented by, for example, Gross et al. (2002). The first step is the *invitation step* where researchers invite a group of individuals from the population of interest. Invitation is observed with the key assumption that the response function is mean independent of invitation. The second step is the *participation step* where invitees decide whether to participate. The experiment is conducted only on the participants, which is the source of the selection problem. The third step is the actual experiment involving random assignment of treatment. Under these assumptions, the width of the identified set relates to the fraction of *invitees* that participate, since invitees are assumed to be representative of the overall population. Without distinguishing between invitation and participation, as in the model of Manski (1996), the width of the identified set relates to the fraction of the *population* that participates. This fraction can be extremely small in practice, and perhaps even unknown by the econometrician. Therefore, that approach to RT data results in very wide, and perhaps even uninformative, identified sets. Therefore, the three step formulation of the model is an important feature of the model, as it distinguishes between the "selection" that is under the control of the experimenter and the "selection" that is under the control of potential subjects. This can

result in tighter identification, because it allows the econometrician to assume that invitation is not selective while still allowing that participation is selective. Hence, the model in this paper and the model in Manski (1996) contain complementary results on identification from RT data under different models and sets of assumptions.

Policy evaluation experiments in economics often have the structure that a population of interest is determined, which are often "units" that can be assigned distinct policy treatments. Then, "units" from that population are invited to participate, "units" decide whether to participate, and some randomization of policy is applied to the participating "units." Such experiments fit the three step model of an RT. For example, the "oversubscription method" as described by Duflo et al. (2007) is one where "demand for a program or service exceeds supply" and the RT "select[s] those who will receive the program by lottery among eligible candidates." In such an RT, invitation could be the "eligibility criteria" and perhaps an actual invitation to apply, and participation could be the attempt to actually sign up for the program. In such an RT, the selectivity of participation relates to the idea that the individuals that actively exhibit a "demand" for a program may be selectively drawn from the overall population of interest. Of course, not all RTs follow exactly this process. Even if the three step model of the RT is not adequate for some specific experiments, the underlying idea of this paper may still apply: use the structure of the experimental design to guide the assumptions made in the identification analysis.

This paper also considers additional assumptions that tighten the identified set, and in particular result in point identification. Specifically, the paper shows that the standard estimator of an average treatment effect from RT data, namely the difference in average outcomes between the groups of participants in the RT that are assigned different treatments, is robust to certain failures of the standard assumption that participation in the RT is not selective. The paper also provides an identification result based on an instrument for participation that makes it possible to test for selective participation.

Other papers in economics also deal with evaluating the benefits of identification based on experimental data, for example Heckman (1992, 1996). Of course, a fundamental difference is the partial identification approach taken in this paper.

In addition, another feature of this paper is the comparison between identification based on RT data and identification based on observational data. This question relates to research design, specifically deciding whether to use RT data or observational data. RT data is subject to selection bias due to selective participation in the RT, and observational data is subject to selection bias due to treatment endogeneity. Hence, there is partial identification of treatment effects from both types of data. Perhaps surprisingly, on the basis of the size of the identified sets, RT data is not necessarily preferred to observational data, due to the fact that the two types of data involve different types of selection (rather than one type of data necessarily having less selection than the other type of data). Observational data is preferred when there are few treatments and/or when there is low participation in the RT.

This paper focuses on identification of the ATR and ATE. These quantities tend to be the focus of identification analyses in causal inference models, because these quantities concern the counterfactual responses to manipulating the treatment. However, other quantities may also be important. For example, if the treatment cannot be directly manipulated, then an intention-to-treat analysis that relates outcomes to *assignment* to treatment may be important. In this model of an RT, it is assumed that there is perfect compliance with treatment assignment, while allowing for possible selectivity of participation in the RT. Therefore, one possible related intention-to-treat analysis could concern the relationship between the outcomes and being *invited to participate* in the RT. This could be relevant, for example, when evaluating the overall "effect" of running the RT. However, this is not pursued in this paper. Estimation and inference can follow by application of existing results for partially identified models, summarized for example in Canay and Shaikh (2017). The identification results are also illustrated using an analysis of RTs in the literature.

Sections 2 and 3 introduce the model and derive the identification results. Because of selective participation in an RT, data from an RT need not point identify the ATR or ATE.

However, under further assumptions an RT can point identify either the ATR or the ATE. One such assumption is that the response function is mean independent of participation, but others are that participation is non-selective conditional on observables or that there is a separable effect of participation on response. These results specifically depend on this particular model of an RT, and would not generically apply to other "one stage" missing data and/or causal inference problems. These sections also consider identification of the sign of the ATE, which is the object of interest when the experimenter wants to learn the "best" treatment and is abstracting away from issues like different costs of different treatments, and identification of whether participation is selective. Section 4 compares identification from RT data to identification from observational data, characterizing when RT data results in a narrower identified set than does observational data, and vice versa. Furthermore, under this model of the RT, it is possible to point identify the sign of the ATE under certain conditions, in contrast to identification with observational data where it is generically not possible to point identify the sign of the ATE under the maintained assumptions. Section 5 illustrates the identification results using an analysis of RTs in the literature. Finally, Section 6 concludes with some suggestions for reporting the results of RTs.

2. Identification in a three stage model of an RT

The finite set of possible treatments is $\mathcal{D}$. Each individual i in the population has a response function $y_i(\cdot) : \mathcal{D} \to \mathbb{R}$. The randomized trial (RT) consists of three decisions at the individual level. The first decision is the decision by the experimenter of whether to invite individual i to participate in the RT. The binary variable indicating an invitation is I_i. Invitation is observed. The second decision is the decision by an invited individual of whether to participate in the RT. The binary variable indicating participation is P_i. Participation is observed. By construction, if $I_i = 0$ then $P_i = 0$. Finally, the third decision is the randomized assignment of a treatment to each subject participating in the RT. The variable indicating assigned treatment is D_i. Treatment is observed. Further, in order to focus on issues concerning external validity due to selection into the RT, the analysis will abstract from many important issues that might threaten internal validity of an RT. In particular,

the analysis abstracts away from issues related to treatment compliance and assumes that all subjects comply with their assigned treatment. The actual outcome of subject i is Y_i. Because of perfect compliance, $Y_i = y_i(d)$ when $D_i = d$. And there are known finite bounds on the possible outcomes, so that necessarily $y_i(d) \in [m, M]$ for all treatments d.

The population information that forms the basis for the identification analysis is the distribution $P(I_i, P_iI_i, D_iP_iI_i, Y_iP_iI_i)$. In other words, for each individual in the population of interest, the econometrician observes: whether that individual is invited to the RT, whether an invited individual (with $I_i = 1$) elects to participate, and the treatment and outcome of an individual that was invited and elects to participate (with $P_i = 1$ and $I_i = 1$).

Therefore, $P_iI_i = 1$ is the condition for information on the treatment and outcome of individual i to be in the RT data. Invitation and participation collectively determine which individuals "actually" participate in the RT. Despite that, it is useful to separate the invitation and participation decisions to make clear the meaning of the assumptions that are used in the analysis of the data, since it is useful to distinguish between the invitation decision that is under the control of the experimenter and the participation decision that is not under the control of the experimenter. This three step model of an RT is similar to the stylized model presented by, for example, Gross et al. (2002).

The identification analysis begins with $E(y_i(d))$, the average response to treatment d. This analysis suppresses any regressors in the analyses for simplicity, but everything can involve non-parametric conditioning on regressors. By the law of iterated expectations,

$$E(y_i(d)) = \underbrace{E(y_i(d)|I_i = 1)}_{(1)} \underbrace{P(I_i = 1)}_{(2)} + \underbrace{E(y_i(d)|I_i = 0)}_{(3)} \underbrace{P(I_i = 0)}_{=1-(2)}.$$

In other words, the average response to treatment d can be decomposed as the weighted average of the average treatment responses for individuals invited, (1), and not invited, (3), to the RT where the weight is the probability of invitation, (2). The first substantiative assumption is that the response function is mean independent of invitation.

Assumption 1. *Assume that* $E(y_i(d)|I_i = 1) = E(y_i(d)) = E(y_i(d)|I_i = 0)$ *for all treatments* $d \in \mathcal{D}$. *Also,* $P(I_i = 1) > 0$.

Of course, an even stronger version of the assumption would involve full statistical indpendence between the response function and invitation, but full statistical independence is not needed for the identification result on average treatment response and average treatment effect. Directly, and without using Assumption 1, the identification analysis establishes sharp bounds for $E(y_i(d)|I_i = 1)$. Under Assumption 1, these bounds immediately imply the same bounds are sharp also for $E(y_i(d))$. In principle, if Assumption 1 were not assumed, but other assumptions were made, sharp bounds on $E(y_i(d))$ would additionally involve an analysis of what the data and the maintained assumptions identify about $E(y_i(d)|I_i = 0)$. The maintained definition in this paper is that the RT data has no information about the response functions for individuals not participating in the RT. Such information could, in principle, come from other data sources other than the RT data. But many RTs study the response to an experimental treatment that is not available to individuals not invited to participate in the RT. In that case, any possible data alone is completely uninformative about $E(y_i(d)|I_i = 0)$ when d is such an experimental treatment, since no individual not participating in the experiment experiences treatment d.

The credibility of Assumption 1 depends on the relationship between the experimental design and the population of interest. It is always possible to define the population of interest to be exactly the invited population, in which case $I_i = 1$ for all individuals. However, in most cases there is a population of interest defined beyond the limited scope of the invitees for a particular RT. In general, the credibility of Assumption 1 depends on how the individuals are invited from the population of interest, and in particular whether they are invited in a way that is correlated with their response function. Critically, this is under the control of the experimenter, unlike the participation decision made by the invited individuals. See also Van Spall et al. (2007) for a more general study of invitation to medical RTs, in particular.

The identification power of this three stage model of an RT, and specifically the identification power of I_i and associated Assumption 1, is to maintain that the subpopulation invited to participate can be taken to be representative of the population of interest. Invitation is under the control of the experimenter. Then, the possibility that participation among the invited is

selective is the focus of this paper. Participation among the invited is not under the control of the experimenter. Without Assumption 1, or in a model that does not distinguish between the invitation step and the participation step, the width of the identified set would relate to the fraction of the *population* that participates in the RT. With Assumption 1, the width of the identified set relates to the fraction of the *invitees* that participates in the RT. The latter fraction can be much greater than the former fraction, resulting in much tighter identification with Assumption 1 compared to without Assumption 1. Indeed, without Assumption 1, or in a model that does not distinguish between the invitation step and the participation step, the result would be an almost completely uninformative identified set, when the number of subjects involved in the RT is very small compared to the size of the population of interest.

Assumption 1 can be satisfied even if there is some overlap between the decision maker(s) responsible for the invitation decision and the decision maker(s) responsible for the participation decision. For example, the "experimenter" may unilaterally make the invitation decisions and also advise the invited individuals on whether to participate, perhaps on the basis of some beliefs about whether participation in the RT will help or hurt the invited individual. Such an RT can still satisfy Assumption 1, since the "experimenter" can make the invitation without consideration of the response function even if the "experimenter" then goes on to advise the individual about participation with consideration of the response function. The previous remarks in this paper about the identification power of Assumption 1 would still apply in this case. For example, in RTs with "ethical considerations" as in the introduction, it is possible that an overlapping set of researchers are involved in invitation and advising on participation.

As motivated by Gross et al. (2002), this model and assumption is an exact characterization of some RTs, and at least a good approximation of some other RTs. However, this model and assumption may not be a good approximation to certain types of RTs. In particular, if the experimenter only has selective access to individuals from the population of interest, then it may not be credible to assume that invitation to the RT is mean independent of the response function as required by Assumption 1, since obviously $I_i = 0$ for individuals that

the experimenter does not even have access to. (Of course, "access to" is ambiguous. The key point is that there may be reasons that the experimenter cannot actually invite individuals from the population in a way that is mean independent of the response function, even if the experimenter "intended to.") In that case, the econometrician can redefine the "population" in the identification analysis to be the population that the experimenter has access to. The rest of the identification analysis can proceed as is, relative to that redefined population. In that case, the identification analysis would recover information about treatment effects on the population that the experimenter has access to, which may not be representative of the overall population. For example, some small scale RTs may only invite from a "convenience population," like college students. Therefore, an important consideration when applying these results is a careful specification of the overall population of interest, in order to evaluate whether individuals indeed are invited without respect to their response function from that population of interest.

Another application of the law of iterated expectations implies that

$$\begin{aligned} E(y_i(d)|I_i=1) = &\underbrace{E(y_i(d)|P_i=1,I_i=1)}_{(1)}\underbrace{P(P_i=1|I_i=1)}_{(2)} \\ &+\underbrace{E(y_i(d)|P_i=0,I_i=1)}_{(3)}\underbrace{P(P_i=0|I_i=1)}_{=1-(2)}. \end{aligned}$$

In other words, the average response to treatment d among those invited to participate can be decomposed as the weighted average of the average treatment response for individuals participating, (1), and not participating, (3), in the RT where the weight is the probability of participation given invitation, (2). As before, the RT data alone is completely uninformative about $E(y_i(d)|P_i=0,I_i=1)$, since the RT data has no information about the response functions for individuals not participating in the RT. The identification analysis maintains this definition of RT data when establishing sharpness of the bounds. The identification analysis also assumes that the RT is "ideal" in the sense that it point identifies $E(y_i(d)|P_i=1,I_i=1)$. A sufficient condition for this is the usual mean independence of the response function from treatment assignment (and perfect compliance) that usually defines an "ideal" RT.

Assumption 2. *Assume that* $E(y_i(d)|P_i = 1, I_i = 1) = E(y_i(d)|D_i = d, P_i = 1, I_i = 1)$. *Assume further that* $P(P_i = 1|I_i = 1) > 0$ *and* $P(D_i = d|P_i = 1, I_i = 1) > 0$ *for all treatments* $d \in \mathcal{D}$.

Under this assumption, $E(y_i(d)|P_i = 1, I_i = 1) = E(Y_i|D_i = d, P_i = 1, I_i = 1)$. In other words, the average response to treatment d is point identified *for subjects actually participating in the experiment*. These two assumptions seem to exhaust the assumptions that can be credibly made on RTs in general. The resulting identified set for $E(y_i(d))$ is given by the following theorem.

Theorem 1. *Under Assumptions 1 and 2, the sharp identified set for* $E(y_i(d))$ *is that*

$$E(y_i(d)) \in E(Y_i|D_i = d, P_i = 1, I_i = 1)P(P_i = 1|I_i = 1) + [m, M]P(P_i = 0|I_i = 1).$$

Further, the sharp identified set for $\{E(y_i(d))\}_{d\in\mathcal{D}}$ *is the Cartesian product of these sets.*

Proof. The previous discussion establishes these bounds. Sharpness is obtained by considering the response functions

$$y_i(d) = \begin{cases} Y_i & \text{if } D_i = d, P_i = 1, I_i = 1 \\ E(Y_i|D_i = d, P_i = 1, I_i = 1) & \text{if } D_i \neq d, P_i = 1, I_i = 1 \\ [m, M] & \text{if } P_i = 0, I_i = 1 \\ E(y_i(d)|I_i = 1) & \text{if } I_i = 0. \end{cases}$$

These response functions are consistent with the data by the first line of the definition, are consistent with Assumption 1 by the fourth line, and are consistent with Assumption 2 by the second line. They also obviously achieve any point in the identified set by the third line. □

Corollary 2. *Under the same assumptions as Theorem 1, the sharp identified set for* $ATE(d', d) \equiv E(y_i(d') - y_i(d))$ *is that*

$$ATE(d', d) \in \Big(E(Y_i|D_i = d', P_i = 1, I_i = 1) - E(Y_i|D_i = d, P_i = 1, I_i = 1)\Big)P(P_i = 1|I_i = 1)$$

$$+ [m - M, M - m]P(P_i = 0|I_i = 1).$$

This corollary follows from the theorem since the identified set for $E(y_i(d')) \times E(y_i(d))$ is the Cartesian product of the marginal identified sets, since there are not restrictions across treatments. The next corollary considers identification of the sign of the ATE, which may be of specific importance in some cases. This is the case when the experimenter wants to learn the "best" treatment and is abstracting away from issues like different costs of different treatments, which would lead to the magnitude of the ATE mattering. Define the experimental ATE as $ATE_{exp}(d', d) \equiv E(Y_i|D_i = d', P_i = 1, I_i = 1) - E(Y_i|D_i = d, P_i = 1, I_i = 1)$.

Corollary 3. *Under the same assumptions as Theorem 1, $ATE(d', d)$ is point identified to be positive (or non-negative, resp.) if $ATE_{exp}(d', d)P(P_i = 1|I_i = 1) + (m - M)P(P_i = 0|I_i = 1) > (\geq)0$, to be negative (or non-positive, resp.) if $ATE_{exp}(d', d)P(P_i = 1|I_i = 1)+(M-m)P(P_i = 0|I_i = 1) < (\leq)0$, and is not identified and can be positive, negative, or zero if $ATE_{exp}(d', d)P(P_i = 1|I_i = 1) + (m - M)P(P_i = 0|I_i = 1) < 0$ and $ATE_{exp}(d', d)P(P_i = 1|I_i = 1) + (M - m)P(P_i = 0|I_i = 1) > 0$.*

The first key conclusion of Theorem 1 is that, under these assumptions, RT data is informative about the average treatment response as long as a positive fraction of invited individuals participate in the experiment. The second key conclusion is that unless the participation in the RT is 100 percent among those invited, i.e. $P(P_i = 1|I_i = 1) = 1$, there is not point identification of the average treatment response. Note from Corollary 3 that even though there is not point identification of the ATE when $P(P_i = 1|I_i = 1) < 1$ there can be point identification of the sign of the ATE in many cases. Basically, the condition for point identification of the sign of the ATE is that the ATE in the subpopulation of individuals participating in the experiment is sufficiently large in magnitude relative to the fraction of individuals participating in the experiment to outweigh any possible ATE in the subpopulation of individuals not participating in the experiment. Note this contrasts with identification with observational data, as in Section 4.

2.1. **Why selective participation in RTs?** The reason that there is not point identification of the ATR and ATE in general is that the identified set accounts for the possibility that the response function is not mean independent of participation in the RT. The reasons for concern that the response function is not mean independent of realized treatment in observational data are basically exactly the same reasons that there could be concern that the response function is not mean independent of participation in an RT. The simple reason is that participation in an RT amounts to a gamble relating to receiving the treatments in the RT.

This claim is consistent with a simple economic theory of how individuals decide whether to participate in an RT. Suppose that individuals (or their agents; for example, their caregivers in a medical setting) have preferences over the treatments they receive. Suppose in particular that the utility is the same as the actual outcomes that result from these treatments. This abstracts away from, for example, differences in the costs of these treatments. It also requires that the individuals perfectly know the outcome that results from each of the treatments. Of course, the goal of the RT is to learn the outcome that results from the treatments, so this assumption is almost certainly not literally true. But, it seems a good first approximation to motivate why the assumption that the response function is mean independent of participation in an RT is not necessarily credible for all RTs.

Suppose that treatments $\mathcal{D}_{ne} \subset \mathcal{D}$ are the non-experimental treatments available outside of the experiment, treatments $\mathcal{D}_e \subset \mathcal{D}$ are the experimental treatments available only in the experiment, and that treatments $\mathcal{D}_{RT} \subset \mathcal{D}$ are the treatments available in the experiment, which might include some non-experimental treatments. By assumption, $\mathcal{D}_e \subset \mathcal{D}_{RT}$. Suppose for simplicity that individuals are risk-neutral expected utility maximizers, with utility equal to the outcome, and that they know the probability they will receive each treatment should they participate in the RT. Then individual i will participate if and only if $\sum_{d \in \mathcal{D}_{RT}} p_d y_i(d) \geq \max_{d \in \mathcal{D}_{ne}} y_i(d)$, where $p_d > 0$ is the probability of receiving treatment d in the RT. A necessary condition is that $\max_{d \in \mathcal{D}_e} y_i(d) \geq \max_{d \in \mathcal{D}_{ne}} y_i(d)$, since otherwise participation in the RT is dominated by not participating and being able to choose the optimal non-experimental

treatment.[1] In particular, suppose that there is exactly one experimental treatment, d_e. Then a necessary condition for participation is that $y_i(d_e) \geq \max_{d \in \mathcal{D}_{ne}} y_i(d)$. Despite being a very simplified model of the participation decision, this suggests that participants in an RT will tend to not be representative of the population of all individuals invited to participate.

Also, Section 3.4 provides an identification result based on an instrument for participation that makes it possible to test for selective participation.

3. Extra assumptions for identification in a three stage model of an RT

The next parts consider identification under additional assumptions.

3.1. Response function is mean independent of participation.

If the response function is mean independent of participation in the RT, there is the following point identification result.

Assumption 3. *Assume that* $E(y_i(d)|P_i = 1, I_i = 1) = E(y_i(d)|I_i = 1) = E(y_i(d)|P_i = 0, I_i = 1)$ *for all treatments* $d \in \mathcal{D}$.

Theorem 4. *Under Assumptions 1, 2, and 3,* $E(y_i(d))$ *is point identified as* $E(Y_i|D_i = d, P_i = 1, I_i = 1)$.

Suppose that the experimenter maintains that Assumptions 1 and 2 hold, and then "assumes" that the average treatment response is point identified as $E(y_i(d)) = E(Y_i|D_i = d, P_i = 1, I_i = 1)$. This is akin to assuming that the decision to participate in the RT among invitees is not selective, in the sense that the response function is mean independent of participation, by the following argument. Recall that under Assumptions 1 and 2

$$\begin{aligned} E(y_i(d)) &\underbrace{=}_{1} E(y_i(d)|I_i = 1) \\ &= E(y_i(d)|P_i = 1, I_i = 1)P(P_i = 1|I_i = 1) \\ &\quad + E(y_i(d)|P_i = 0, I_i = 1)P(P_i = 0|I_i = 1) \end{aligned}$$

[1] Let $y_i^* = \max_{d \in \mathcal{D}_{ne}} y_i(d)$. Then re-write $\sum_{d \in \mathcal{D}_{RT}} p_d y_i(d) \geq \max_{d \in \mathcal{D}_{ne}} y_i(d)$ as $\sum_{d \in \mathcal{D}_e} p_d(y_i(d) - y_i^*) + \sum_{d \in \mathcal{D}_{RT} \cap \mathcal{D}_{ne}} p_d(y_i(d) - y_i^*) \geq 0$. The second sum is non-positive by definition of y_i^*, so $\sum_{d \in \mathcal{D}_e} p_d(y_i(d) - y_i^*) \geq 0$ and consequently $\max_{d \in \mathcal{D}_e} y_i(d) \geq y_i^*$.

$$\underbrace{=}_{2} E(Y_i|D_i = d, P_i = 1, I_i = 1)P(P_i = 1|I_i = 1)$$
$$+ E(y_i(d)|P_i = 0, I_i = 1)P(P_i = 0|I_i = 1).$$

Therefore, by algebra, for the condition $E(y_i(d)) = E(Y_i|D_i = d, P_i = 1, I_i = 1)$ to hold it must be that either $P(P_i = 1|I_i = 1) = 1$, which seems non-generic, or $E(y_i(d)|P_i = 0, I_i = 1) = E(Y_i|D_i = d, P_i = 1, I_i = 1) = E(y_i(d)|P_i = 1, I_i = 1)$. But this is precisely Assumption 3. So under Assumptions 1 and 2 the conventional estimate of the ATR used in an RT is equivalent to the assumption that the response function is mean independent of participation in the RT. Section 3.5 considers related questions for the conventional estimate of the ATE.

More generally, it is sufficient for the conventional interpretation of an RT as point identifying the average treatment response that three mean independence assumptions hold: the response function is mean independent of invitation, participation, and treatment assignment. An ideal RT should satisfy the first and third assumptions, but not necessarily the second assumption. The first and third assumptions are under the control of the experimenter but the second assumption is not. Again, this is the motivation for the model that distinguishes between these aspects of the RT data.

Remark 1 (Justifying the assumptions)**.** It is often the case that experimental studies report summary statistics that are used to suggest that randomization has "worked" because the observables of the subjects receiving each treatment (including perhaps the subjects in the control group when applicable) have similar distributional properties. Note that while this may bolster the case for Assumption 2, it implies nothing about whether Assumptions 1 and/or 3 are true, which concern whether the response function is mean independent of invitation/participation. A useful measure of whether the response function is mean independent of invitation/participation would be a comparison of the same distributional properties between invitees and non-invitees, and between participants and non-participants among the invitees; depending on the nature of these covariates, this comparison may or may not be feasible. For example, if the measurement of the covariates is invasive and non-standard then almost by definition of non-participation, individuals who do not participate will have

missing data for these covariates. It is possible, however, to consider combining many different datasets. It is also important to note even this cannot "prove" that the response function is mean independent of invitation/participation because of the possibility of unobservables. Section 5 includes further discussion on this issue.

3.2. **Participation is non-selective conditional on observables.** If there is a suitable observable such that participation is non-selective conditional on the observable, it is possible to point identify the average treatment response. This observable random variable X_i is discrete (for simplicity), and is characterized by the following assumption.

Assumption 4. *Assume that X_i is an observed variable for all individuals who are invited to participate and that $E(y_i(d)|X_i = x, P_i = 1, I_i = 1) = E(y_i(d)|X_i = x, I_i = 1) = E(y_i(d)|X_i = x, P_i = 0, I_i = 1)$ for all treatments $d \in \mathcal{D}$ and all x in the support of $X|(I = 1)$. Also assume that $P(P_i = 1|I_i = 1) > 0$. And assume also that $P(X_i = x|P_i = 1, I_i = 1) > 0$ and $P(X_i = x|P_i = 0, I_i = 1) > 0$ for all x in the support of $X|(I = 1)$.*

In other words, the response function is mean independent of participation conditional on the control. The analysis also assumes that the RT is "ideal" in the sense that the response function is mean independent of treatment assignment conditional now also on the control.

Assumption 5. *Assume that $E(y_i(d)|X_i = x, P_i = 1, I_i = 1) = E(y_i(d)|D_i = d, X_i = x, P_i = 1, I_i = 1)$ and further that $P(D_i = d|X_i = x, P_i = 1, I_i = 1) > 0$ for all treatments $d \in \mathcal{D}$ and all x in the support of $X|(I = 1)$.*

Then

$$\begin{aligned} E(y_i(d)|P_i = 1, I_i = 1) &= \sum_x E(y_i(d)|X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 1, I_i = 1) \\ &= \sum_x E(Y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 1, I_i = 1) \end{aligned}$$

and similarly

$$E(y_i(d)|P_i = 0, I_i = 1) = \sum_x E(y_i(d)|X_i = x, P_i = 0, I_i = 1)P(X_i = x|P_i = 0, I_i = 1)$$

$$
\begin{aligned}
&= \sum_x E(y_i(d)|X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 0, I_i = 1) \\
&= \sum_x E(Y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 0, I_i = 1).
\end{aligned}
$$

This establishes the following theorem.

Theorem 5. *Under Assumptions 1, 4, and 5, $E(y_i(d))$ is point identified as $\sum_x E(Y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|I_i = 1)$.*

Proof. Using the prior expressions,

$$
\begin{aligned}
&E(y_i(d)|I_i = 1) \\
&= E(y_i(d)|P_i = 1, I_i = 1)P(P_i = 1|I_i = 1) + E(y_i(d)|P_i = 0, I_i = 1)P(P_i = 0|I_i = 1) \\
&= \left(\sum_x E(Y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 1, I_i = 1)\right) P(P_i = 1|I_i = 1) \\
&+ \left(\sum_x E(Y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|P_i = 0, I_i = 1)\right) P(P_i = 0|I_i = 1) \\
&= \left(\sum_x E(Y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x, P_i = 1|I_i = 1)\right) \\
&+ \left(\sum_x E(Y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x, P_i = 0|I_i = 1)\right) \\
&= \left(\sum_x E(Y_i|D_i = d, X_i = x, P_i = 1, I_i = 1)P(X_i = x|I_i = 1)\right)
\end{aligned}
$$

□

In other words, the average response to treatment d is the weighted average of the average outcomes for subjects participating in the experiment and assigned treatment d who have covariates x, weighted by the probability that an individual invited to participate has covariates x. In the special case that $P(X_i = x|D_i = d, P_i = 1, I_i = 1) = P(X_i = x|I_i = 1)$ so that the distribution of the control is independent of participating in the experiment (and receiving treatment d) then this simplifies to the usual RT result that $E(y_i(d)) = E(Y_i|D_i = d, P_i = 1, I_i = 1)$ like in Theorem 4. Otherwise, identification of the average treatment response needs to involve re-weighting by the distribution of the controls among all

individuals invited to participate, not the distribution of the controls among those actually participating.

3.3. **Participation is selective conditional on excluded instrument.** Section 3.2 requires the existence of observables such that participation is non-selective conditional on observables. In some settings, a more credible identification strategy may rest on the existence of an instrument such that the instrument has no direct effect on the response function, while allowing that participation is possibly selective conditional on the instrument. If the instrument shifts the probability of participation in the RT, an identification strategy that uses the instrument can result in a tighter identified set.

Specifically, suppose that Z_i is an instrumental variable that satisfies the standard conditions for being an instrument for participation in the RT:

Assumption 6. *Assume that Z_i is an observed variable for all individuals who are invited to participate and that $E(y_i(d)|Z_i = z, I_i = 1) = E(y_i(d)|I_i = 1)$ for all treatments $d \in \mathcal{D}$ and all z in the support of $Z|(I = 1)$.*

Note that this implicitly involves the assumption of an exclusion restriction, in the sense that the instrument is "excluded" from affecting the average response to the treatment. The analysis also assumes that the RT is "ideal" in the sense that the response function is mean independent of treatment assignment conditional now also on the instrument.

Assumption 7. *Assume that $E(y_i(d)|Z_i = z, P_i = 1, I_i = 1) = E(y_i(d)|D_i = d, Z_i = z, P_i = 1, I_i = 1)$ and further that $P(D_i = d|Z_i = z, P_i = 1, I_i = 1) > 0$ for all treatments $d \in \mathcal{D}$ and all z in the support of $Z|(I = 1)$.*

Under those assumptions, for any z in the support of $Z|(I = 1)$, it follows that

$$\begin{aligned} E(y_i(d)) &\underbrace{=}_{1} E(y_i(d)|I_i = 1) \\ &\underbrace{=}_{6} E(y_i(d)|Z_i = z, I_i = 1) \\ &= E(y_i(d)|Z_i = z, P_i = 1, I_i = 1)P(P_i = 1|Z_i = z, I_i = 1) \end{aligned}$$

$$+ E(y_i(d)|Z_i = z, P_i = 0, I_i = 1)P(P_i = 0|Z_i = z, I_i = 1)$$

$$\underbrace{=}_{7} E(Y_i|D_i = d, Z_i = z, P_i = 1, I_i = 1)P(P_i = 1|Z_i = z, I_i = 1)$$

$$+ E(y_i(d)|Z_i = z, P_i = 0, I_i = 1)P(P_i = 0|Z_i = z, I_i = 1).$$

Theorem 6. *Under Assumptions 1, 6, and 7, the sharp identified set for $E(y_i(d))$ is that*

$$E(y_i(d)) \in \cap_z \{E(Y_i|D_i = d, Z_i = z, P_i = 1, I_i = 1)P(P_i = 1|Z_i = z, I_i = 1) + [m, M]P(P_i = 0|Z_i = z, I_i = 1)\}.$$

Further, the sharp identified set for $\{E(y_i(d))\}_{d\in\mathcal{D}}$ is the Cartesian product of these sets.

Proof. The previous discussion establishes these bounds. Sharpness is obtained by considering the response functions

$$y_i(d) = \begin{cases} Y_i & \text{if } D_i = d, P_i = 1, I_i = 1 \\ E(Y_i|D_i = d, Z_i = z, P_i = 1, I_i = 1) & \text{if } D_i \neq d, Z_i = z, P_i = 1, I_i = 1 \\ t_{zd} & \text{if } Z_i = z, P_i = 0, I_i = 1 \\ E(y_i(d)|I_i = 1) & \text{if } I_i = 0, \end{cases}$$

where $t_{zd} = \frac{\psi_d - E(Y_i|D_i=d,Z_i=z,P_i=1,I_i=1)P(P_i=1|Z_i=z,I_i=1)}{P(P_i=0|Z_i=z,I_i=1)}$, for a given $\psi_d \in \cap_z \{E(Y_i|D_i = d, Z_i = z, P_i = 1, I_i = 1)P(P_i = 1|Z_i = z, I_i = 1) + [m, M]P(P_i = 0|Z_i = z, I_i = 1)\}$.

These response functions are consistent with the data by the first line of the definition, are consistent with Assumption 1 by the fourth line, are consistent with Assumption 7 by the second line, and consistent with Assumption 6 by the first, second, and third lines. They also obviously achieve any point in the identified set by the third line. □

3.4. **Testing for selective participation.** It is also possible to point identify the average treatment responses (and therefore average treatment effects) for the subpopulation of individuals whose participation decision is manipulatable by the instrument. This is similar to the local average treatment effect of Imbens and Angrist (1994), except the instrument is for participation rather than the treatment. This provides scope for testing for selective participation in the RT.

Let $\mathcal{Z}$ be the support of the instrument, which is assumed to be an ordered set. Also suppose that each individual i in the population has a participation decision function $p_i(\cdot) : \mathcal{Z} \to \{0, 1\}$ that describes the binary participation decision of individual i as a function of the value of the instrument assigned to individual i. The assumptions needed for this result are essentially the same as used in Imbens and Angrist (1994), except for participation rather than the treatment, and adapted to account for the assumption of random assignment of treatment in the RT.

Assumption 8. *The participation decision function $p_i(z)$ is a weakly increasing function of z, for all individuals i. The instrument for participation and the random assignment of treatment in the RT satisfy the condition that $((y_i(\cdot), p_i(\cdot)) \perp (Z_i, D_i))|(I_i = 1)$. The invitation decision similarly satisfies the condition that $(y_i(\cdot), p_i(\cdot)) \perp I_i$. Further, $P(I_i = 1) > 0$ and $P(D_i = d|Z_i = z, P_i = 1, I_i = 1) > 0$ for all treatments $d \in \mathcal{D}$ and all z in the support of $Z|(I = 1)$.*

This assumption basically subsumes Assumptions 1, 6, and 7. Note that it requires that treatment assignment be defined for all invitees, even invitees that do not actually participate in the experiment.

Theorem 7. *Under Assumption 8, for any $z \in \mathcal{Z}$ and $z' \in \mathcal{Z}$ with $z' > z$,*

$$E(y_i(d)|p_i(z') = 1, p_i(z) = 0) = \frac{E(P_i Y_i|D_i = d, Z_i = z', I_i = 1) - E(P_i Y_i|D_i = d, Z_i = z, I_i = 1)}{E(P_i|Z_i = z', I_i = 1) - E(P_i|Z_i = z, I_i = 1)},$$

as long as the denominator is non-zero.

Proof. By definition of the response function and participation decision function, and then by two applications of Assumption 8, $E(P_i Y_i|D_i = d, Z_i = z, I_i = 1) = E(p_i(z)y_i(d)|D_i = d, Z_i = z, I_i = 1) = E(p_i(z)y_i(d)|I_i = 1) = E(p_i(z)y_i(d))$. Therefore, $E(P_i Y_i|D_i = d, Z_i = z', I_i = 1) - E(P_i Y_i|D_i = d, Z_i = z, I_i = 1) = E((p_i(z') - p_i(z))y_i(d))$. Because $p_i(\cdot)$ is weakly increasing by Assumption 8, and $z' > z$, $E((p_i(z') - p_i(z))y_i(d)) = E(y_i(d)|p_i(z') = 1, p_i(z) = 0)P(p_i(z') = 1, p_i(z) = 0)$. By similar arguments, $E(P_i|Z_i = z, I_i = 1) = E(p_i(z)|Z_i =$

$z, I_i = 1) = E(p_i(z)|I_i = 1) = E(p_i(z))$, so $E(P_i|Z_i = z', I_i = 1) - E(P_i|Z_i = z, I_i = 1) = E(p_i(z') - p_i(z)) = P(p_i(z') = 1, p_i(z) = 0)$. □

This result makes it possible to test for selective participation in the RT. If $E(y_i(d)|p_i(z') = 1, p_i(z) = 0) \neq E(y_i(d)|p_i(z''') = 1, p_i(z'') = 0)$, for some $z' > z$ and $z''' > z''$, then that is evidence for selective participation in the RT, since it means that the participation decision function is related to the response function. More specifically, it means that individuals that are induced to participate in the RT at different levels of the instrument experience different responses to the treatment d. For example, the instrumental variable could be a monetary incentive that an individual receives for participating in the RT. If the response to the treatment is different for individuals induced to participate at different levels of the monetary incentives, then that is evidence for selective participation in the RT.

This test is related to the fact that overidentification tests in standard instrumental variables models are implicitly tests of treatment effect homogeneity. However, this test concerns selectivity of participation ("endogeneity" of participation). The difference is because the instrument in this setup is an instrument for participation, rather than an instrument for the treatment as in the standard instrumental variables setup.

3.5. **A separable participation effect.** Under a functional form assumption on the response function, even if the response function is not mean independent of participation in the RT it is possible to point identify the average treatment effect by "differencing out" the effect of participation.

Suppose that $y_i(d) = y_{i0}(d) + \alpha_i$. Under a suitable assumption, α_i captures all selectivity of participation in the RT. Then, $y_i(d') - y_i(d) = y_{i0}(d') - y_{i0}(d)$, so the treatment effect does not depend on α_i. The key identifying assumption is as follows.

Assumption 9. *There exists $y_{i0}(\cdot)$ and α_i with $y_i(d) = y_{i0}(d) + \alpha_i$ for all $d \in \mathcal{D}$ and all individuals, and such that $E(y_{i0}(d)|P_i = 1, I_i = 1) = E(y_{i0}(d)|I_i = 1) = E(y_{i0}(d)|P_i = 0, I_i = 1)$ for all $d \in \mathcal{D}$.*

In other words, α_i captures all of the selectivity of participation in the RT. The restriction imposed by this assumption is that any selectivity of participation affects the response to all treatments equally. Then under the assumption that the response function is mean independent of treatment assignment, Assumption 2, $E(Y_i|D_i = d, P_i = 1, I_i = 1) = E(y_i(d)|P_i = 1, I_i = 1) = E(y_{i0}(d)|P_i = 1, I_i = 1) + E(\alpha_i|P_i = 1, I_i = 1) = E(y_{i0}(d)|I_i = 1) + E(\alpha_i|P_i = 1, I_i = 1)$. Consequently, under the assumption that the response function is mean independent of invitation, Assumption 1, $ATE(d', d) \equiv E(y_i(d') - y_i(d)) = E(y_i(d') - y_i(d)|I_i = 1)$ is point identified by $E(Y_i|D_i = d', P_i = 1, I_i = 1) - E(Y_i|D_i = d, P_i = 1, I_i = 1) = E(y_{i0}(d')|I_i = 1) - E(y_{i0}(d)|I_i = 1) = E(y_i(d') - y_i(d)|I_i = 1)$. This establishes the following theorem.

Theorem 8. *Under Assumptions 1, 2, and 9, the average treatment effect* $ATE(d', d) \equiv E(y_i(d') - y_i(d))$ *is point identified as* $ATE_{exp}(d', d) \equiv E(Y_i|D_i = d', P_i = 1, I_i = 1) - E(Y_i|D_i = d, P_i = 1, I_i = 1)$. *However, the sharp identified set for* $E(y_i(d))$, *for* any one *treatment* d, *remains the same as in Theorem 1.*

Proof. The identification of $ATE(d', d)$ follows from the previous discussion. The result that the sharp identified set for $E(y_i(d))$ remains the same as in Theorem 1 is obtained by considering the response functions

$$y_{i0}(d) = \begin{cases} Y_i & \text{if } D_i = d, P_i = 1, I_i = 1 \\ E(Y_i|D_i = d, P_i = 1, I_i = 1) & \text{if } D_i \neq d, P_i = 1, I_i = 1 \\ E(Y_i|D_i = d, P_i = 1, I_i = 1) & \text{if } P_i = 0, I_i = 1 \\ E(y_i(d)|I_i = 1) & \text{if } I_i = 0. \end{cases}$$

and

$$\alpha_i = \begin{cases} 0 & \text{if } D_i = d, P_i = 1, I_i = 1 \\ 0 & \text{if } D_i \neq d, P_i = 1, I_i = 1 \\ [m, M] - E(Y_i | D_i = d, P_i = 1, I_i = 1) & \text{if } P_i = 0, I_i = 1 \\ 0 & \text{if } I_i = 0. \end{cases}$$

These response functions $y_{i0}(\cdot)$ and α_i add up to the same response function used to establish sharpness in the proof of Theorem 1. Therefore they are consistent with the data, and Assumptions 1 and 2, and achieve any point in the identified set. They also satisfy Assumption 9. □

It is useful that $ATE_{exp}(d', d)$ identifies the average treatment effect *either* under the standard assumption that participation is not selective (Assumption 3) or under the alternative assumption that participation is selective but has an additively separable effect on the response function (Assumption 9). Consequently, the standard estimate of the treatment effect in an RT is automatically robust to certain failures of the assumption that participation is not selective.

The result in this section places a functional form assumption on the response function. If instead the participation effect enters the response function non-separably, differencing as in this section may not remove the participation effect from the treatment effect. In some cases, a non-separable effect of participation on response is plausible. For example, it could be that the participation effect captures characteristics of the subjects that interact differently with the different treatments. For example, it could be that participation captures the "motivation" of the subject and that the response to some treatments depends on motivation while the response to other treatments does not depend on motivation.

4. Comparing RT data to observational data

This section compares identification based on RT data to identification based on observational data. Specifically, this section considers the identification of the model with RT data as

derived in Theorem 1. Therefore, this section compares the "no assumptions" identification with RT data to the "no assumptions" identification with observational data.

Theorem 9 (Manski (2007))**.** *Under no assumptions, with observational data the identified set for* $E(y_i(d))$ *is* $E(y_i(d)) \in E(Y_i|D_i = d)P(D_i = d) + [m, M]P(D_i \neq d)$.

Further, the identified set for $ATE(d', d)$ *is* $ATE(d', d) \in E(Y_i|D_i = d')P(D_i = d') - E(Y_i|D_i = d)P(D_i = d) + [mP(D_i \neq d') - MP(D_i \neq d), MP(D_i \neq d') - mP(D_i \neq d)]$.

4.1. **On the basis of identifying the sign of the ATE.** The lower bound on the identified set for $ATE(d', d)$ with observational data can be written as $(E(Y_i|D_i = d') - M)P(D_i = d') + (m - E(Y_i|D_i = d))P(D_i = d) + (m - M)(1 - P(D_i = d') - P(D_i = d))$. The upper bound can be written as $(E(Y_i|D_i = d') - m)P(D_i = d') + (M - E(Y_i|D_i = d))P(D_i = d) + (M - m)(1 - P(D_i = d') - P(D_i = d))$.

If $m < M$, which is generic since otherwise there is no variation in outcomes, these expressions make clear that the sign of $ATE(d', d)$ is completely unidentified with observational data except possibly in the case that $1 - P(D_i = d') - P(D_i = d) = 0$. Otherwise, for the lower bound the first two terms are non-positive and the last term is negative, and similarly for the upper bound the first two terms are non-negative and the last term is positive. Therefore both strictly positive and strictly negative values for ATE are in the identified set, so the sign of $ATE(d', d)$ is completely unidentified. In case $1 - P(D_i = d') - P(D_i = d) = 0$ and $P(D_i = d) > 0$ and $P(D_i = d') > 0$ then a non-negative ATE is point identified exactly in case $E(Y_i|D_i = d') = M$ and $E(Y_i|D_i = d) = m$. A non-positive ATE is point identified exactly in case $E(Y_i|D_i = d') = m$ and $E(Y_i|D_i = d) = M$. It is never the case that a zero ATE can be ruled out with observational data, since it is always consistent with the data that $y_i(d') = Y_i = y_i(d)$ for all d', d.

On the other hand, recall from Corollary 3 that the sign of the ATE can be point identified with RT data. Therefore, for an RT satisfying the conditions of Theorem 1, and specifically the conditions of Corollary 3, RT data can have better identification power than does observational data for the purposes of identifying the sign of the ATE.

4.2. **On the basis of the widths of the identified sets.** The width of the identified set for the average treatment response $E(y_i(d))$ with RT data is $(M-m)P(P_i=0|I_i=1)$ while with observational data it is $(M-m)P(D_i \neq d)$. The width depends on the treatment considered with observational data but does not with RT data. Therefore, comparison of identification with RT data and identification with observational data depends on the participation rate, and the treatment considered and the details of how treatments are selected in the observational data.

In order to compare the two types of data in a general way, consider the sum of the widths of identified sets across all treatments, which is a measure of the total "ambiguity" (lack of knowledge) that remains about the average treatment responses. With RT data this is $(M-m)|\mathcal{D}|P(P_i=0|I_i=1)$. With observational data this is $\sum_{d\in\mathcal{D}}(M-m)(1-P(D_i=d)) = (M-m)|\mathcal{D}|(1-\frac{1}{|\mathcal{D}|})$.

Therefore, the RT data is better than the observational data if and only if $P(P_i=0|I_i=1) < 1-\frac{1}{|\mathcal{D}|}$ or equivalently $|\mathcal{D}|P(P_i=1|I_i=1) > 1$. That is, the RT data is preferred when there are many treatments and/or when there is high participation in the RT among those invited.

Conversely, the observational data is better than the RT data if and only if $|\mathcal{D}|P(P_i=1|I_i=1) < 1$. That is, the observational data is preferred when there are few treatments and/or when there is low participation in the RT among those invited. The intuition for this result is that while RT data point identifies the ATR on the subpopulation of participants, it could be that there is highly selective participation. It is possible, even though observational data has the usual treatment selection problem, that the observational data has information on a greater fraction of the population of interest, compared to the fraction of participants in the RT, in which case the observational data may result in narrower identified sets.

A similar result obtains for the sum of the widths of the identified sets for the average treatment effects. The identified sets for $ATE(d',d)$ is the difference in the identified sets for $E(y_i(d'))$ and $E(y_i(d))$ for RT data and for observational data since there are no assumptions that involve restrictions across treatments. Therefore the width of the identified set for

$ATE(d', d)$ is the sum of the widths of the identified sets for $E(y_i(d'))$ and $E(y_i(d))$. Let $\mathcal{H}(\cdot)$ be the identified set for its argument, and let $|\mathcal{H}(\cdot)|$ be the width of the identified set. Therefore $\sum_{d\in\mathcal{D}}\sum_{d'>d}|\mathcal{H}(ATE(d',d))| = \sum_{d\in\mathcal{D}}\sum_{d'>d}(|\mathcal{H}(E(y_i(d')))| + |\mathcal{H}(E(y_i(d)))|) = (|\mathcal{D}|-1)\sum_{d\in\mathcal{D}}|\mathcal{H}(E(y_i(d)))|$.

Therefore the comparison between RT data and observational data on the basis of the sums of the widths of the identified sets for the average treatment effects is the same as the comparison on the basis of the sums of the widths of the identified sets for the average treatment responses.

Remark 2 (Combining RT and observational data)**.** This paper has derived various bounds on treatment response using RT data. If the econometrician has access to RT data and other (observational) data, then bounds can be combined. A simple way to do that is to obtain bounds on, for example, the ATR using both RT data and observational data and then forming the intersection of these bounds to get an overall bound on ATR.

5. Empirical illustration

This section illustrates the identification results using some facts about randomized trials from the literature. In this illustration, it is supposed that the outcome is binary. Consequently, $m = 0$ and $M = 1$.[2] Table 1 shows how the participation rate among those invited, $P(P_i = 1|I_i = 1)$, relates to the identification of the sign of the average treatment effect. It is supposed that there are two treatments of interest, and that some RT satisfying the conditions of Theorem 1 reveals the experimental ATE, $ATE_{exp}(d', d) \equiv E(Y_i|D_i = d', P_i = 1, I_i = 1) - E(Y_i|D_i = d, P_i = 1, I_i = 1)$. Recall this is the ATE on the subpopulation that actually participates in the experiment. The table shows the smallest experimental ATE such that according to Corollary 3 the ATE in the population is point identified to be non-negative.

[2]The additional assumption of a discrete outcome rather than a continuous outcome does not affect the sharpness of the bounds. In proving sharpness, the proof exhibited response functions that are the same for all people in certain subpopulations (e.g., the response to treatment d is the same for everybody with $I_i = 0$). These exhibited response functions in general will not take values compatible with discreteness of the outcome, but necessarily the response functions take values in the convex hull of the set of outcomes. It is trivial to simply partition any given subpopulation further and assign people in that sub-subpopulation to have outcomes compatible with the discreteness of the outcome, and such that on average that subpopulation has the same outcome as does the subpopulation in the sharpness proofs.

If the participation rate is too low it is never possible to point identify the population ATE to be non-negative, and this is indicated in the table by "n.p." for not possible.

Each row of the table provides a possible participation rate among the invited and the corresponding smallest experimental ATE that point identifies the population ATE to be non-negative. So if the participation rate among invited is 60%, for example, then the experimental ATE must be at least as great as $\frac{2}{3}$ in order to point identify the population ATE to be non-negative. If the participation rate is strictly less than 50% then it is not possible to point identify the population ATE to be non-negative, because even if the experimental ATE were 1, the largest possible, the ATE in the subpopulation that does not participate but is invited could be -1, which would result in a negative population ATE. Note that the marginal gain in identifying power from increasing the participation rate is greatest when the participation rate among the invited is low, in the sense that the derivative of the smallest ATE_{exp} implying the population ATE is non-negative is (when it exists) $-P(P_i = 1|I_i = 1)^{-2}$, and so is decreasing in magnitude in the participation rate among the invited. Consequently, there is relatively less gain from increasing participation among the invited from 90% to 100% and relatively more gain from increasing participation among the invited from 50% to 60%. For contrast, with observational data, per Section 4.1, it is generically not possible to point identify the sign of the ATE.

Participation rate among invited	Smallest ATE_{exp} implying $ATE \geq 0$
10%	n.p.
20%	n.p.
30%	n.p.
40%	n.p.
50%	1
60%	$\frac{2}{3} \approx .67$
70%	$\frac{3}{7} \approx .43$
80%	$\frac{1}{4} = .25$
90%	$\frac{1}{9} \approx .11$
100%	0

TABLE 1. Effect of participation rate on identification of the sign of the average treatment effect; n.p. = not possible, there is no ATE_{exp} implying $ATE \geq 0$.

For context, consider an analysis of recruitment into medical RTs in Gross et al. (2002). Gross et al. (2002) study 172 medical RTs published over the course of a year in four major medical journals. In these RTs the median eligibility fraction, the fraction of the potential participants who are eligible to enroll in the study after screening (roughly analogous to invitation in the model in Section 2), is 65%. The interquartile range is 41-82. These figures are based on the 48 studies that report the necessary data in the publication. The median enrollment fraction, the fraction of eligible participants who actually enroll (roughly analogous to participation among the invited in the model in Section 2), is 93%. The interquartile range is 79-100. These figures are based on the 74 studies that report the necessary data.[3] The median recruitment fraction, the product of these two fractions, is 54%. The interquartile range is 32-77. These figures are based on 81 studies that report the necessary data. With this enrollment fraction (participation rate among invited), a modest experimental average treatment effect is sufficient to point identify that the population average treatment effect is non-negative. Therefore, in such studies, even without any assumptions about selectivity of participation, it may be realistically possible to point identify the existence of a positive treatment effect, as long as the treatment indeed does have a positive effect. Similar conclusions could be drawn when Banerjee and Duflo (2009) report an experiment of a "no legal strings attached" gift worth "between $25 and $100" as part of the Bandhan microfinance program in India. Approximately 18% of the invited individuals (translating roughly to the model) rejected the gift, a participation rate similar to the above.

In other RTs, the participation rate is even lower. One such example is the experimental study of the Job Training Participation Act (JTPA) conducted by the Manpower Demonstration Research Corporation. In this experiment, Doolittle and Traeger (1990) and Heckman (1992) report that more than 90% of invited training centers refused to participate. With a participation rate of less than 10%, such data cannot even identify the sign of the average

[3]Note that this implies less than one-half of studies report this data. It is not obvious which direction the resulting median is biased from the median in the population. It could be that studies with a low enrollment fraction are more likely to report that in the research paper, because it may threaten the validity of the study and so it is worth reporting in the research paper. Alternatively, it could be that studies with a low enrollment fraction are less likely to get published in a major medical journal, exactly because a low enrollment fraction may threaten validity.

treatment effect, unless the econometrician is willing to assume something about selectivity of participation.

This analysis of the identification of the sign of the average treatment effect involves no assumptions on selectivity of participation, but if the econometrician does assume that participation is not selective, then the experimental average treatment effect is the population average treatment effect (Theorem 4). Therefore, a key question is whether it is credible to assume that participation is not selective, and the participation rate by itself does not necessarily imply anything about selectivity of participation. In many cases by definition of not participating in the RT limited data is available on individuals who do not participate. Nevertheless, it may be possible to compare the characteristics of subjects who participate with, for example, population data from other sources, in order to get a sense of whether participation seems likely to be related to the response function. This sort of analysis is conducted by Steg et al. (2007) in a meta-analysis of RTs of treatments for acute myocardial infarction ("heart attack"). They find that the characteristics of patients who are eligible for an RT but do not participate are "worse" than of patients who actually participate. The same pattern holds for the observed outcomes.[4] Rothwell (2005) provides results that suggest that even this sort of comparison may not be enough because of characteristics that are likely unobserved in the data (at the time of the recruitment into the RT) but are related to participation. Rothwell (2005) reports that in an RT of endarterectomy to prevent stroke that roughly 3% of the patients were randomized into receiving the endarterectomy but "did not have surgery because their surgeon and/or anaesthetist judged them to be too frail." This group had a distribution of observables (that appear in the dataset) similar to that for the rest of the participants in the RT but did have a much higher subsequent rate of stroke compared to the patients participating in the RT but not receiving the endarterectomy. This suggests that showing that observable characteristics of participants in an RT are similar to those not participating may not be enough to establish that the response function is mean independent of participation, depending on the nature of the observables.

[4]Their analysis also suggests that invitation in these RTs may not satisfy Assumption 1.

Similarly in the case of the Job Training Participation Act (JTPA) experiment there is suggestive evidence that participation by training centers is selective. Relevant data is reported by Doolittle and Traeger (1990, Tables 5.4-5.5). The training centers that participated tended to, among other things, be geographically un-representative of all centers, serve a smaller number of terminees from Title IIA of the JPTA, and had greater adult employment rates.

Some of these examples depend on covariates, and if there are observed covariates that suitably explain participation then the strategy of Section 3.2 can be used.

6. Conclusions

This paper studies the question of what is identified about the average treatment response (ATR) and average treatment effect (ATE) with data from a randomized trial. The paper focused on the problem of selective participation in an RT.

The analysis of this paper suggests that in reporting the results of an RT it is useful to consider reporting the bounds on the ATR and ATE as derived in this paper. If this is not possible then it is useful to report information to the extent possible on the invitation rate; how individuals are invited to participate; the characteristics of those who are invited, and are not invited; the participation rate of those invited; and the characteristics of those who participate, and are invited but do not participate. This view is partly seen in the CONSORT statement (i.e., Moher et al. (2010) and Schulz et al. (2010)), a major set of guidelines for medical RTs, which states that "[a] comprehensive description of the eligibility criteria used to select the trial participants is needed to help readers interpret the study" and "[a] description of the method of recruitment, such as by referral or self selection (for example, through advertisements), is also important in this context."

It is also worth noting that, for the purposes of identification, it is the participation *rate* that matters, not the absolute number of participants. This suggests that the emphasis in designing RTs should be on high participation rates, not simply a large number of participants. This is because participants in a large RT can be equally or less representative of the population of interest than participants in a small RT if the two RTs differ in their emphasis on recruitment of a representative subpopulation. Of course, statistical precision is improved with a larger

sample. But a statistically precise estimate of a less informative identified set may be less preferred than an imprecise estimate of a more informative identified set.

References

Banerjee, A. V. and E. Duflo (2009). The experimental approach to development economics. *Annual Review of Economics 1*(1), 151–178.

Canay, I. A. and A. M. Shaikh (2017). Practical and theoretical advances in inference for partially identified models. In B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson (Eds.), *Advances in Economics and Econometrics: Eleventh World Congress*, Volume 2 of *Econometric Society Monographs*, pp. 271–306. Cambridge University Press.

Doolittle, F. and L. Traeger (1990). *Implementing the National JTPA Study.*

Duflo, E., R. Glennerster, and M. Kremer (2007). Using randomization in development economics research: A toolkit. In T. P. Schultz and J. A. Strauss (Eds.), *Handbook of Development Economics*, Volume 4, Chapter 61, pp. 3895–3962. Elsevier.

Gross, C., R. Mallory, A. Heiat, and H. Krumholz (2002). Reporting the recruitment process in clinical trials: who are these patients and how did they get there? *Annals of Internal Medicine 137*(1), 10–16.

Heckman, J. J. (1992). Randomization and social policy evaluation. In C. F. Manski and I. Garfinkel (Eds.), *Evaluating Welfare and Training Programs*, Chapter 5. Harvard University Press.

Heckman, J. J. (1996). Randomization as an instrumental variable. *The Review of Economics and Statistics 78*(2), 336–341.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Manski, C. F. (1996). Learning about treatment effects from experiments with random assignment of treatments. *The Journal of Human Resources 31*(4), 709–733.

Manski, C. F. (2007). *Identification for Prediction and Decision.* Harvard University Press.

Moher, D., S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ 340*, c869.

Rothwell, P. (2005). External validity of randomised controlled trials: To whom do the results of this trial apply? *The Lancet 365*(9453), 82–93.

Schulz, K. F., D. G. Altman, and D. Moher (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ 340*, c332.

Steg, P., J. Lopez-Sendon, E. Lopez de Sa, S. Goodman, J. Gore, F. Anderson Jr, D. Himbert, J. Allegrone, and F. Van de Werf (2007). External validity of clinical trials in acute myocardial infarction. *Archives of Internal Medicine 167*(1), 68–73.

Van Spall, H. G. C., A. Toren, A. Kiss, and R. A. Fowler (2007). Eligibility criteria of randomized controlled trials published in high-impact general medical journals: A systematic sampling review. *JAMA 297*(11), 1233–1240.

Weijer, C., M. W. Enkin, S. H. Shapiro, and K. C. Glass (2000). Clinical equipoise and not the uncertainty principle is the moral underpinning of the randomised controlled trial. *BMJ 321*(7263), 756–758.

World Medical Association (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA 310*(20), 2191–2194.