# MIT Open Access Articles

## *Experimental designs for identifying causal mechanisms*

# Experimental Designs for Identifying Causal Mechanisms[*]

Kosuke Imai[†]        Dustin Tingley[‡]        Teppei Yamamoto[§]

Forthcoming in *Journal of the Royal Statistical Society, Series A*
*(with discussions)*

## Abstract

Experimentation is a powerful methodology that enables scientists to empirically establish causal claims. However, one important criticism is that experiments merely provide a black-box view of causality and fail to identify causal mechanisms. Specifically, critics argue that although experiments can identify average causal effects, they cannot explain the process through which such effects come about. If true, this represents a serious limitation of experimentation, especially for social and medical science research that strive to identify causal mechanisms. In this paper, we consider several different experimental designs that help identify average natural indirect effects. Some of these designs require the perfect manipulation of an intermediate variable, while others can be used even when only imperfect manipulation is possible. We use recent social science experiments to illustrate the key ideas that underlie each of the proposed designs.

**Key Words:** causal inference, direct and indirect effects, identification, instrumental variables, mediation

[†]Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

[‡]Assistant Professor, Government Department, Harvard University. Email: dtingley@princeton.edu, URL: http://scholar.harvard.edu/dtingley

[§]Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: teppei@mit.edu, URL: http://web.mit.edu/teppei/www

# 1 Introduction

Over the last century and across numerous disciplines, experimentation has been a powerful methodology to test scientific theories. As Fisher (1935) demonstrated a long time ago, the key advantage of randomized experiments is their ability to estimate causal effects without bias. However, one important criticism is that experiments merely provide a black-box view of causality. Many critics have argued that while experiments can identify average causal effects, they cannot explain causal mechanisms (e.g., Heckman and Smith, 1995; Cook, 2002; Deaton, 2009). If true, this represents a serious limitation of experimentation, especially for social and medical science research which strives to identify how treatments work.

In this paper, we study how to design randomized experiments to identify causal mechanisms. We use the term causal mechanism to mean a causal process through which the effect of a treatment on an outcome comes about. This is motivated by the fact that many applied researchers, especially those in social sciences, use the term to refer to such a process. We formalize the concept of causal mechanism by what is known in the methodological literature as "natural indirect effect," or "causal mediation effect," which quantifies the extent to which the treatment affects the outcome through the mediator (e.g., Robins and Greenland, 1992; Pearl, 2001; Imai *et al.*, 2010b, see Section 2 for more discussion).

To identify causal mechanisms, the most common approach taken by applied researchers is what we call the *single experiment design* where causal mediation analysis is applied to a standard randomized experiment. This approach is popular in psychology and other disciplines (e.g., Baron and Kenny, 1986). However, as formally shown by many researchers, it requires strong and untestable assumptions that are similar to those made in observational studies. Thus, the use of the single experiment design is often difficult to justify from an experimentalist's point of view.

To overcome this problem, we propose alternative experimental designs. First, in Section 3, we consider two designs useful in situations where researchers can directly manipulate the intermediate variable that lies on the causal path from the treatment to the outcome. Such a variable is often referred to as a "mediator" and we follow this convention throughout this paper. Under the *parallel design*, each subject is randomly assigned to one of two experiments; in one experiment only the treatment variable is randomized while in the other both the treatment and the mediator are randomized. Under the *crossover design*, each experimental unit is sequentially assigned to two experiments where the first assignment is conducted randomly and the subsequent assignment is determined without randomization based on the treatment and mediator values in the previous experiment. We show that the two designs have a potential to significantly

improve the identification power of the single experiment design.

Despite their improved identification power, the parallel and crossover designs have disadvantages that are not shared by the single experiment design. First, it is often difficult to perfectly manipulate the mediator. For example, in psychological experiments, the typical mediators of interest include emotion and cognition. Second, even if such a manipulation is possible, the use of these designs requires the consistency assumption that the manipulation of mediator should not affect the outcome through any pathway other than the mediator. In medical and social science experiments with human subjects, this often implies that experimental subjects need to be kept unaware of the manipulation. This consistency assumption may be difficult to satisfy especially if manipulating the mediator requires an explicit intervention.

To address these limitations, in Section 4, we propose two new experimental designs that can be used in the situations where the manipulation of the mediator is not perfect (see Mattei and Mealli, 2011, for a related experimental design). These designs permit the use of indirect and subtle manipulation, thereby potentially enhancing the credibility of the required consistency assumption. Under the *parallel encouragement design*, experimental subjects who are assigned to the second experiment are randomly encouraged to take (rather than assigned to) certain values of the mediator after the treatment is randomized. Similarly, the *crossover encouragement design* employs randomized encouragement rather than the direct manipulation in the second experiment. Therefore, these two designs generalize the parallel and crossover designs by allowing for imperfect manipulation. We show that under these designs one can make informative inferences about causal mechanisms by focusing on a subset of the population.

Throughout the paper, we use recent experimental studies from social sciences to highlight key ideas behind each design. These examples are used to illustrate how applied researchers may implement the proposed experimental designs in their own empirical work. In Section 5, we use a numerical example based on actual experimental data to illustrate our analytical results. Section 6 gives concluding remarks.

## 2   The Fundamental Problem of Identifying Causal Mechanisms

In this section, we argue that what many applied researchers mean by "causal mechanisms" can be formalized (and quantified) by using the concepts of direct and indirect effects (Robins and Greenland, 1992; Pearl, 2001). We then briefly discuss the fundamental problem that arises when identifying causal mechanisms. We also discuss an alternative definition of causal mechanisms which focuses on causal components instead of processes (e.g. Rothman, 1976; VanderWeele and Robins, 2009), as well as other related quantities that have appeared in recent works (Geneletti, 2007; Spencer *et al.*, 2005).
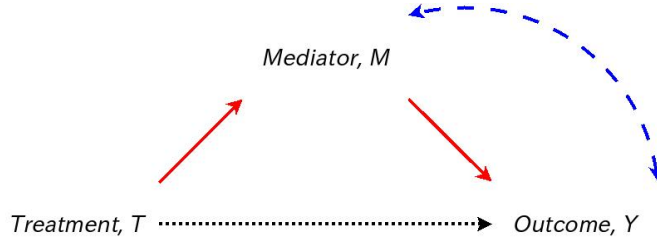
Figure 1: Diagram for a Simple Causal Mechanism. The red solid arrows represent a causal mechanism of interest where the causal effect of the treatment on the outcome is transmitted through the intermediate variable or the mediator. The black dotted arrow represents all the other possible causal mechanisms. The blue dashed arc indicates the possible presence of confounders between the mediator and outcome, which typically cannot be ruled out in the standard single experiment design.

## 2.1 Causal Mechanisms as Direct and Indirect Effects

In this paper, we use the term causal mechanisms to represent the process through which the treatment causally affects the outcome. This viewpoint is widespread throughout social sciences and also is consistent with a common usage of the term in a variety of scientific disciplines (e.g. Salmon, 1984; Little, 1990). Specifically, we study the identification of a simple causal mechanism, which is represented by the red solid arrows in the causal diagram presented in Figure 1. In this diagram, the causal effect of the treatment $T$ on the outcome $Y$ is transmitted through an intermediate variable or a mediator $M$. The black dotted arrow represents all the other possible causal mechanisms of the treatment effect. Thus, the treatment effect is decomposed into the sum of the *indirect effect* (a particular mechanism through the mediator of interest) and the *direct effect* (which includes all other possible mechanisms). From this point of view, identifying the role of the mediator corresponding to the causal pathway of interest allows researchers to learn about the causal process through which a particular treatment affects an outcome.

To formally define indirect effects within the potential outcomes framework, consider a randomized experiment where $n$ units are randomized into the treatment group $T_i = 1$ or the control group $T_i = 0$. Let $M_i \in \mathcal{M}$ denote the observed value of the mediator that is realized after the exposure to the treatment where $\mathcal{M}$ is the support of $M_i$. Since the mediator can be affected by the treatment, there exist two potential values, $M_i(1)$ and $M_i(0)$, of which only one will be observed, i.e., $M_i = M_i(T_i)$. Next, let $Y_i(t, m)$ denote the potential outcome that would result if the treatment variable and the mediator equal $t$ and $m$, respectively. Again, we only observe one of the potential outcomes, i.e., $Y_i = Y_i(T_i, M_i(T_i))$. Throughout this paper, we assume no interference among units; i.e., the potential outcomes of one unit do not depend on the values of the treatment variable and the mediator of another unit (Cox, 1958). We also assume for the sake of simplicity that the treatment variable is binary (i.e., $T_i \in \{0, 1\}$) for the rest of the

paper. Extension to non-binary treatments is possible but beyond the scope of this paper.

Given this setup, the (total) causal effect of the treatment for each unit can be defined as,

$$\tau_i \quad \equiv \quad Y_i(1, M_i(1)) - Y_i(0, M_i(0)). \tag{1}$$

Now, the unit indirect effect at the treatment level $t$ is defined as,

$$\delta_i(t) \quad \equiv \quad Y_i(t, M_i(1)) - Y_i(t, M_i(0)), \tag{2}$$

for $t = 0, 1$ (Robins and Greenland, 1992; Pearl, 2001; Robins, 2003). The key to understanding equation (2) is the following counterfactual question: what change would occur to the outcome if one changes the mediator from the value that would realize under the control condition, i.e., $M_i(0)$, to the value that would be observed under the treatment condition, i.e., $M_i(1)$, while holding the treatment status at $t$? Because these two values of the mediator are the ones that would naturally occur as responses to changes in the treatment, the quantity in equation (2) formalizes the notion of a causal mechanism that the causal effect of the treatment is transmitted through changes in the mediator of interest.

Similarly, we define the unit direct effect, corresponding to all other possible causal mechanisms, as,

$$\zeta_i(t) \quad \equiv \quad Y_i(1, M_i(t)) - Y_i(0, M_i(t)), \tag{3}$$

for $t = 0, 1$. The key counterfactual question is: what difference in the outcome would result if one changes the treatment status from $T_i = 0$ to $T_i = 1$ while holding the mediator value constant at $M_i(t)$? In some cases (see Section 3.2), the direct effect rather than the indirect effect is of interest to scientists.

According to Holland (1986), the fundamental problem of causal inference under the potential outcomes framework is that given any unit one cannot observe the potential outcomes under the treatment and control conditions at the same time. The problem of identifying causal mechanisms is more severe than that of identifying causal effects. In particular, while $Y_i(t, M_i(t))$ is observable for units with $T_i = t$, $Y_i(t, M_i(1 - t))$ is *never* observed for any unit regardless of its treatment status without additional assumptions. This implies that although it identifies the average treatment effect $\bar{\tau}$, the randomization of the treatment alone can neither identify the average indirect effect $\bar{\delta}(t)$ nor the average direct effect $\bar{\zeta}(t)$. These average effects are defined as follows, $\bar{\tau} \equiv \mathbb{E}(Y_i(1, M_i(1)) - Y_i(0, M_i(0)))$, $\bar{\delta}(t) \equiv \mathbb{E}(Y_i(t, M_i(1)) - Y_i(t, M_i(0)))$, $\bar{\zeta}(t) \equiv \mathbb{E}(Y_i(1, M_i(t)) - Y_i(0, M_i(t)))$, for $t = 0, 1$.

All together, the average indirect and direct effects sum up to the average total effect, i.e., $\bar{\tau} = \bar{\delta}(t) + \bar{\zeta}(1 - t)$. The direct and indirect effects under different treatment status, i.e., $t$ and $1 - t$, need to be combined in order for their sum to equal the total effect. The equality simplifies to $\bar{\tau} = \bar{\delta} + \bar{\zeta}$ when

$\bar{\delta} = \bar{\delta}(1) = \bar{\delta}(0)$ and $\bar{\zeta} = \bar{\zeta}(1) = \bar{\zeta}(0)$. Clearly, these relationships also hold among the unit level effects. Finally, the fact that we can decompose the average causal effect, $\bar{\tau}$, into the sum of average direct and indirect effects, i.e., $\zeta(t)$ and $\delta(t)$, implies that the identification of the average direct effect implies that of the average indirect effect (or vice versa) so long as the average causal effect is also identified. Finally, in Appendix A.1, we briefly discuss a related quantity that appears in the recent work of Geneletti (2007).

## 2.2 Alternative Definitions of Causal Mechanisms

As depicted in Figure 1, we use the term "causal mechanism" to refer to a causal *process* through which the treatment affects the outcome of interest. Clearly, this is not the only definition of causal mechanisms (see Hedström and Ylikoski, 2010, for various definitions of causal mechanisms, many of which are not mentioned here). For example, some researchers define a causal mechanism as a set of *components* which, if jointly present, always produce a particular outcome. This perspective, which can be seen in early works on causality such as Mackie (1965) and Rothman (1976), has been recently formalized under the sufficient cause framework (e.g., Rothman and Greenland, 2005; VanderWeele and Robins, 2007). VanderWeele (2009) formally studies the relationship of this alternative definition to the process-based definition of the causal mechanism using a diagrammatic approach.

Instead of attempting to identify a complete set of sufficient causes, applied researchers often focus on the more tractable task of identifying *causal interactions*. The goal is to test whether or not an outcome occurs only when a certain set of variables are present. To identify causal interactions, the most common practice is to establish statistical interactions between two variables of interest by including their interaction term in a regression model. **?** derive the conditions under which this procedure is justified.

Although justifiable for analyzing causal components, such a procedure is generally not useful for the study of causal processes. For example, while causal interactions between treatment and mediator can be identified by randomizing both variables, such manipulation is not sufficient for the identification of causal processes. To see this formally, note that the existence of a causal interaction between the treatment and the mediator can be defined as,

$$Y_i(1, m) - Y_i(1, m') \quad \neq \quad Y_i(0, m) - Y_i(0, m'), \tag{4}$$

for some $m \neq m'$. This definition makes it clear that the causal interaction exists when the causal effect of a direct manipulation of the mediator varies as a function of the treatment, but not necessarily when the effect of the treatment is transmitted through the mediator. This implies that the non-zero interaction effect per se does not imply the existence of a relevant causal process. In fact, as shown in later sections, under

some experimental designs one must assume the *absence* of interactions to identify causal processes.

Finally, some advocate the alternative definitions of direct and indirect effects based on principal strat-ification (Rubin, 2004) and develop new experimental designs to identify them (Mattei and Mealli, 2011). In this framework, for those units whose mediating variable is not influenced by the treatment at all, the entire treatment effect can be interpreted as the direct effect. However, for the other units, direct and in-direct effects cannot be defined, which makes it difficult to answer the main question of causal mediation analysis, i.e., whether or not the treatment affects the outcome through the mediator of interest (see Van-derWeele, 2008, for further discussions). Thus, in this paper, we focus on the direct and indirect effects as defined in Section 2.1.

## 2.3 Identification Power of the Single Experiment Design

Given the setup reviewed above, we study the *single experiment design*, which is the most common exper-imental design employed by applied researchers in order to identify causal mechanisms. Under the single experiment design, researchers conduct a single experiment where the treatment is randomized. After the manipulation of the treatment, the values of the mediator and then the outcome are observed for each unit.

**Setup.** The randomization of the treatment (possibly conditional on a vector of observed pre-treatment variables $X_i$ as in matched-pair designs) implies that there is no observed or unobserved confounder of the causal relationship between the treatment and the mediator. Figure 1 encodes this assumption since no dashed bidirectional arrow is depicted between $T$ and $M$ or $T$ and $Y$. Formally, this can be written as,

ASSUMPTION 1 (RANDOMIZATION OF TREATMENT ASSIGNMENT)

$$\{Y_i(t', m), M_i(t) \ : \ t, t' \in \{0, 1\}, \ m \in \mathcal{M}\} \ \perp\!\!\!\perp \ T_i \mid D_i = 0$$

*where it is also assumed that* $0 < \Pr(T_i = t \mid D_i = 0)$ *for all* $t$.

Here, $D_i = 0$ represents that unit $i$ belongs to the standard randomized experiment where the treatment is randomized (but the mediator is not). We have introduced this additional notation for later purposes.

**Identification.** As mentioned in Section 2, the randomization of the treatment alone cannot identify causal mechanisms. Thus, for the identification of direct and indirect effects, researchers must rely on an additional assumption which cannot be justified solely by the experimental design. Imai *et al.* (2010b) show that one possible such assumption is that the observed mediator values are conditionally independent of potential outcomes given the actual treatment status and observed pretreatment variables, as if those mediator values were randomly chosen. This assumption can be written formally as,

ASSUMPTION 2 (SEQUENTIAL IGNORABILITY OF THE MEDIATOR) *For $t, t' = 0, 1$, and all $x \in \mathcal{X}$,*

$$Y_i(t', m) \;\perp\!\!\!\perp\; M_i \mid T_i = t, X_i = x, D_i = 0,$$

*where it is also assumed that* $0 < p(M_i = m \mid T_i = t, X_i = x, D_i = 0)$ *for* $t = 0, 1$ *and for all* $m \in \mathcal{M}$.

Here, we explicitly include the vector of pretreatment confounders $X_i$ in the conditioning set because the experimental design does not guarantee the conditional independence between potential outcomes and the observed mediator given the treatment status alone. It can be shown that under this additional assumption, the average indirect effects are nonparametrically identified (see Theorem 1 of Imai *et al.*, 2010b). Under a linearity assumption, this assumption also justifies the common method popularized by Baron and Kenny (1986) (see Imai *et al.*, 2010a). The discussion of other assumptions closely related to Assumption 2 (such as those of Pearl (2001), Robins (2003), and Petersen *et al.* (2006)) can be found in the literature (e.g., Shpitser and VanderWeele, 2011).

In practice, however, many experimentalists find such an identification assumption difficult to justify for the same reason as the unconfoundedness assumption about treatment assignment in observational studies is considered problematic (e.g., Bullock *et al.*, 2010). For example, Assumption 2 is violated if there exist unobserved confounders that affect both the mediator and the outcome. Imai *et al.* (2010b) also point out that while observed *pretreatment* confounders of the relationship between the mediator and outcome can be controlled for in straightforward ways, the mediator-outcome confounders that are *posttreatment* cannot be accommodated even when they are known and observed. These possibilities imply that making Assumption 2 often involves speculation about unobserved characteristics of units and thus may not be desirable from the experimentalists' point of view.

**Sharp bounds.** How important is an additional assumption such as Assumption 2 for the identification of causal mechanisms under the single experiment design? To answer this question, we derive the sharp bounds on the average indirect effects under Assumption 1 alone (see Sjölander, 2009; Kaufman *et al.*, 2009, for the sharp bounds on the average direct effects). These large-sample bounds represent the ranges within which the true values of the average indirect effects are guaranteed to be located (Manski, 1995). For the sake of illustration, we assume both the outcome and mediator are binary. Then, it is straightforward to obtain the sharp bounds using the linear programming approach (Balke and Pearl, 1997).

The expressions for the bounds, which are given in Appendix A.2, imply that the bounds could be shorter than the original bounds (before conducting an experiment, i.e., $[-1, 1]$), but unfortunately they always contain zero and thus are uninformative about the sign of the average indirect effects. This implies that the single experiment design can never provide sufficient information for researchers to know the

direction of the indirect effects without additional assumptions which are not directly justifiable by the experimental design itself. Given the importance of such untestable identification assumptions, some propose to conduct a sensitivity analysis to formally evaluate how robust one's conclusions are in the presence of possible violations of a key identifying assumption (see Imai *et al.*, 2010a,b).

**Example.**   Brader *et al.* (2008) examine how media framing affects citizens' preferences about immigration policy by prompting emotional reactions. In the experiment, subjects first read a short news story about immigration where both the ethnicity of an immigrant and the tone of the story were randomly manipulated in the $2 \times 2$ factorial design. For the ethnicity manipulation, an image of a Latino male and that of a Caucasian male were used. After reading the story, subjects completed a standard battery of survey questions, which measured the mediating variables that comprise a subject's level of anxiety. Respondents were then asked whether immigration should be decreased or increased, which served as the outcome variable of interest.

The primary hypothesis of the original study is that the media framing may influence public opinion by changing the level of anxiety. Specifically, subjects who are assigned to the Latino image and the negative tone would be more likely to oppose immigration and this opposition would be caused through an increasing level of anxiety. The authors found that respondents in the treatment condition (Latino image and negative tone) exhibited the highest levels of anxiety and opposition to immigration. They applied a linear structural equation model to estimate the average indirect effect of the negative Latino frame on policy preferences through changes in anxiety.

Under this single experiment design, only the treatment was randomized. This makes Assumption 2 unlikely to hold, decreasing the credibility of causal mediation analysis. In many psychological experiments including this one, researchers are interested in psychological mechanisms that explain behavioral or attitudinal responses to experimental manipulations. Thus, the mediator of interest is typically a psychological factor that is difficult to manipulate. As a consequence, the single experiment design is frequently used and causal mediation analysis is conducted under a strong assumption of sequential ignorability.

## 3   Experimental Designs with Direct Manipulation

Many critics of the single experiment design view the randomization of mediator as the solution to the identification of causal mechanisms. For example, a popular strategy is the so-called "causal chain" approach where researchers first establish the existence of the causal effect of the treatment on the mediating variable in a standard randomized trial (e.g., Spencer *et al.*, 2005; Ludwig *et al.*, 2011). Then, in a second

(separate) experiment, the mediator is manipulated and its effect on the outcome variable is estimated, which establishes the causal chain linking the treatment and outcome variables. While intuitively appealing, this two-step procedure generally fails to identify the causal process of how the treatment affects the outcome through the mediator. For example, unless the causal effect of the treatment on the mediator and that of the mediator on the outcome are homogenous across units, one can easily construct a hypothetical population for which the average indirect effect is negative even though both the average causal effect of the treatment on the mediator and that of the mediator on the outcome are both positive (e.g., Imai *et al.*, 2011b).

Manipulating the mediator thus does not provide a general solution to the problem of identifying causal mechanisms. This, however, by no means implies that experimental manipulations of the mediator are useless. Here, we consider two experimental designs that may be applicable when the mediator can be directly manipulated.

## 3.1 The Parallel Design

We first consider the *parallel design* in which two randomized experiments are conducted in parallel. Specifically, we randomly split the sample into two experiments. The first experiment is identical to the experiment described in Section 2.3 where only the treatment is randomized. In the second experiment, we simultaneously randomize the treatment and the mediator, followed by the measurement of the outcome variable. In the causal inference literature, Pearl (2001, Theorem 1) implicitly considers an identification strategy under such a design. Our identification analysis differs from Pearl's in that he considers identification under a sequential ignorability assumption similar to Assumption 2. We also derive the sharp bounds on the average indirect effects in the absence of any assumption not justified by the design itself.

**Setup.** Suppose we use $D_i = 0$ ($D_i = 1$) to indicate that unit $i$ belongs to the first (second) experiment. Then, the potential outcome can be written as a function of the experimental design as well as the treatment and the mediator, i.e., $Y_i(t, m, d)$. Because our interest is in identifying a causal mechanism through which the effect of the treatment is naturally transmitted to the outcome, researchers must assume that the manipulation of the mediator in the second experiment itself has no direct effect on the outcome. Specifically, an experimental subject is assumed to reveal the same value of the outcome variable if the treatment and the mediator take a particular set of values, whether or not the value of the mediator is chosen by the subject ($D_i = 0$) or assigned by the experimenter ($D_i = 1$).

Formally, this assumption can be stated as the following consistency assumption,

ASSUMPTION 3 (CONSISTENCY UNDER THE PARALLEL DESIGN) *For all $t = 0, 1$ and $m \in \mathcal{M}$,*

$$Y_i(t, M_i(t), 0) = Y_i(t, m, 1) \quad \text{if} \quad M_i(t) = m,$$

Under this assumption, we can write $Y_i(t, m, d)$ simply as $Y_i(t, m)$ for any $t, m$ and $d$. The importance of Assumption 3 cannot be overstated. Without it, the second experiment provides no information about causal mechanisms (although the average causal effect of manipulating the mediator under each treatment status is identified). If this assumption cannot be maintained, then it is difficult to learn about causal mechanisms by manipulating the mediator.

Since the treatment is randomized in the first experiment, Assumption 1 is guaranteed to hold. Similarly, in the second experiment, both the treatment and the mediator are randomized and hence the following assumption holds under Assumption 3,

ASSUMPTION 4 (RANDOMIZATION OF TREATMENT AND MEDIATOR) *For $t = 0, 1$ and all $m \in \mathcal{M}$,*

$$Y_i(t, m) \perp\!\!\!\perp \{T_i, M_i\} \mid D_i = 1.$$

**Identification.** Unfortunately, Assumptions 1, 3 and 4 alone cannot identify causal mechanisms under the parallel design. To see this formally, note that we can identify $\mathbb{E}(Y_i(t, M_i(t)))$ and $\mathbb{E}(M_i(t))$ from the first experiment and $\mathbb{E}(Y_i(t, m))$ from the second experiment. On the other hand, $\mathbb{E}(Y_i(t, M_i(t')))$ is not identified as the following decomposition shows,

$$\mathbb{E}(Y_i(t, M_i(t'))) = \int \mathbb{E}(Y_i(t, m) \mid M_i(t') = m) \, dF_{M_i | T_i = t', D_i = 0}(m), \tag{5}$$

where $F(\cdot)$ represents the distribution function. The problem is that the first term in the integral, and therefore the left-hand side, cannot be identified unless $Y_i(t, m)$ is independent of $M_i(t')$ (Pearl, 2001, Theorem 1). Furthermore, if the range of the outcome variable is $(-\infty, \infty)$, then this design provides *no* information about the average causal mediation effect without an additional assumption because the left-hand side of equation (5) is also unbounded.

To achieve identification, we may rely upon the additional assumption that there exists no causal interaction between the treatment and the mediator. Using the definition of interaction given in Section 2.2, the assumption can be formalized under Assumption 3 as follows,

ASSUMPTION 5 (NO INTERACTION EFFECT) *For all $m, m' \in \mathcal{M}$ such that $m \neq m'$,*

$$Y_i(1, m) - Y_i(1, m') = Y_i(0, m) - Y_i(0, m').$$

An equivalent assumption was first introduced by Holland (1988) as additivity and later revisited by Robins (2003) for the identification of indirect effects. This assumption implies that the indirect effect depends only on the value of the mediator not the treatment. Note that this assumption must hold for each unit, not just in expectation, the point to which we return shortly below.

The next theorem shows that if one is willing to assume no interaction effect, we can identify causal mechanisms under the parallel design.

THEOREM 1 (IDENTIFICATION UNDER THE PARALLEL DESIGN) *Under Assumptions 1, 3, 4and 5, for $t = 0, 1$, the average indirect effects are identified and given by,*

$$
\begin{aligned}
\bar{\delta}(t) \;=\;& \mathbb{E}(Y_i \mid T_i = 1, D_i = 0) - \mathbb{E}(Y_i \mid T_i = 0, D_i = 0) \\
& - \int \left\{ \mathbb{E}(Y_i \mid T_i = 1, M_i = m, D_i = 1) - \mathbb{E}(Y_i \mid T_i = 0, M_i = m, D_i = 1) \right\} dF_{M_i \mid D_i = 1}(m).
\end{aligned}
$$

Our proof, which closely follows that of Robins (2003), is given in Appendix A.3. Note that the no-interaction assumption leads to $\bar{\delta}(1) = \bar{\delta}(0)$, thereby giving only one expression for both quantities. The theorem implies that in the situations where Assumptions 1, 3, 4, and 5 are plausible, researchers can consistently estimate the average indirect effect by combining the two experiments.

The estimation can proceed in two steps. First, the first term of the expression in Theorem 1 is the average treatment effect on the outcome and can be estimated by calculating the average differences in the outcomes between the treatment and control groups in the first experiment. Next, the second term is the average direct effect of the treatment on the outcome, which is also the average controlled direct effect under Assumption 5 (Robins, 2003). This can be estimated using the information from the second experiment, by computing the differences in the mean outcomes between the treatment and control groups for each value of the mediator, and then averaging these values over the observed distribution of the mediator. Note that Theorem 1 holds regardless of whether the mediator and outcome are continuous or discrete. Our results also allow for any mediator and outcome variable type unless otherwise stated.

It is important to emphasize that the formula given in Theorem 1 generally cannot be used unless the no interaction assumption holds at the unit level (Assumption 5). For illustration, consider the following hypothetical population, which consists of two types of individuals with equal proportions (i.e., 0.5): $M_i(t) = Y_i(t, 1) = Y_i(t', 0) = p$ and $M_i(t') = Y_i(t, 0) = Y_i(t', 1) = 1 - p$ where $p$ takes the value of either 0 or 1. The no interaction assumption is *on average* satisfied for this population because $\mathbb{E}(Y_i(t, 1) - Y_i(t, 0)) = \mathbb{E}(Y_i(t', 1) - Y_i(t', 0)) = 0$. Computing the average indirect effect based on Theorem 1, however, will lead to a severely biased estimate: the estimate converges to $\bar{\delta}(t) = 0$ while the true value is $\bar{\delta}(t) = 1$. The bias arises from the fact that Assumption 5 itself is violated for *any* individual in

this hypothetical population. That is, the average indirect effect depends upon the baseline value of the treatment since $Y_i(t, 1) - Y_i(t, 0) \neq Y_i(t', 1) - Y_i(t', 0)$ for all $i$ in this example.

Unfortunately, Assumption 5 cannot be directly tested since for each unit we only observe one of the four potential outcomes that consist of the assumed equality. However, researchers can test an implication of this assumption by investigating whether the equality holds in expectation, given the fact that $\mathbb{E}(Y_i(t, m))$ is identified in the second experiment. One way to make Assumption 5 credible is to collect pre-treatment characteristics that are known to be related to the magnitude of interaction effects and implement the parallel design within each stratum defined by these pre-treatment variables. Alternatively, a sensitivity analysis such as the one developed by **?** can be used to examine the robustness of empirical findings to the violation of this assumption.

**Sharp bounds.** The importance of the no-interaction assumption can be understood by deriving the sharp bounds on the average indirect effects without this additional assumption. We also compare the resulting bounds with those obtained in Section 2.3 to examine the improved identification power of the parallel design over the single experiment design. In Appendix A.4, we show the formal expressions for the bounds under Assumptions 1, 3 and 4 (but without Assumption 5) for the case in which both the mediator and the outcome are binary. As expected, these bounds are at least as informative as the bounds under the single experiment design because the first experiment under the parallel design gives identical information as the single experiment design as a whole, and the second experiment provides additional information. Moreover, the bounds imply that unlike the single experiment design, the parallel design can sometimes identify the sign of the average indirect effects. However, there exists a tradeoff between the informativeness of the lower bound and that of the upper bound, in that the lower and upper bounds tend to covary positively for both $\bar{\delta}(1)$ and $\bar{\delta}(0)$. This means that the width of the bounds tends to be relatively wide even when the sign of the true value is identified from the data to be either positive or negative.

**Example.** In behavioral neuroscience, scholars have used brain imaging technology, such as functional magnetic resonance imaging (fMRI), to measure the operation of neural mechanisms. fMRI measures local changes in blood flow to particular regions of the brain, which is a proxy for brain activity. Another technology, transcranial magnetic stimulation (TMS), uses repeated magnetic pulses to localized portions of the brain in order to manipulate activation of the region. This allows in principle for a direct manipulation of the hypothesized neural mechanism linking a stimulus with a behavioral response. A growing number of studies use TMS (e.g. Martin and Gotts, 2005; Paus, 2005) because it "directly leads to causal inferences about brain functioning rather than the purely associational evidence provided by imaging tech-

niques" (Camerer *et al.*, 2005, pp.13–14).

For example, Knoch *et al.* (2006) used TMS to understand the neural mechanisms underlying a common behavior observed in the strategic situation known as the "ultimatum game." In this game, a "proposer" decides on the division of a resource worth $R$ by offering $p$ to a "receiver" and keeping $R - p$ for herself. The receiver can either accept the offer (he receives $p$) or reject the offer (both parties receive 0). Standard economic models fail to predict the rejection of positive offers in this game, but it frequently happens in laboratory experiments. One prominent explanation is based on the concept of fairness; individuals tend to reject unfair offers even though their acceptance will be profitable.

In a previous study, Sanfey *et al.* (2003) found fMRI-based evidence that two regions of the brain are activated when subjects decide whether or not to reject an unfair offer: the anterior insula and dorsolateral prefrontal cortex (DLPFC). Given this result, Knoch *et al.* used TMS to investigate whether the activity in DLPFC controls an impulse to reject unfair offers or regulates a selfish impulse. It was argued that if the DLPFC were to be deactivated and individuals accept more unfair offers, then this would represent evidence that the DLPFC serves the role of implementing fair behavior and regulating selfish impulses, instead of inhibiting fairness impulses.

The parallel design may be applicable in this setting. Here, the treatment variable is whether an individual receives a fair or unfair offer, and thus can be easily randomized. With the aid of TMS, researchers can also directly manipulate the mediator by changing the activity level of the DLPFC. The outcome variable, whether or not the offer was rejected, can then be measured. As discussed above, the key identification assumption is the consistency assumption (Assumption 3), which mandates in this context that subjects must not be aware of the fact that they were being manipulated. In the original study, every subject wore the same TMS apparatus, and none of them were aware of whether or not they were actually exposed to the stimulation by the TMS, increasing the credibility of the consistency assumption. However, such manipulation may be difficult in practice even with technologies such as TMS, because anatomical localization for TMS device placement is known to be imperfect (Robertson *et al.*, 2003).

For the parallel design, the no interaction effect assumption is required for the identification of causal mechanisms. Is this assumption reasonable in the Knoch *et al.* experiment? Their results suggest not. They find that the effect of changing the mediator in the fair offers condition is less than in the unfair offers condition. Although this result can be taken as evidence that the fairness of offers and the activation of the DLPFC causally interact in determining subjects' behavior, this does not necessarily imply that the DLPFC represents a causal process through which the effect of the fairness treatment is transmitted.

## 3.2 The Crossover Design

To further improve on the parallel design, we must directly address the fundamental problem of identifying causal mechanisms discussed in Section 2. For example, we can never observe $M_i(1)$ for units with $T_i = 0$, but we must identify $\mathbb{E}(Y_i(0, M_i(1)))$ in order to identify $\bar{\delta}(0)$. Here, we consider the *crossover design* where each experimental unit is exposed to both the treatment and control conditions sequentially. This design differs from the standard crossover designs in an important way (Jones and Kenward, 2003). Specifically, under this design, the experimenter first randomizes the order in which each unit is assigned to the treatment and control conditions. After measuring the value of the mediator and then that of the outcome variable, each unit is assigned to the treatment status opposite to their original treatment condition and to the value of the mediator that was observed in the first period. Optionally, the second stage of this design can be modified to include a randomly-selected subgroup for each treatment group which does not receive the mediator manipulation (see below for the rationale behind this possible modification). Finally, the outcome variable is observed for each unit at the end of the second period.

The intuition behind the crossover design is straightforward; if there is no carryover effect (as defined formally below), then the two observations for each unit can be used together to identify the required counterfactual quantities. This design is different from the one suggested by Robins and Greenland (1992) where "both exposure and the cofactor intervention [i.e. mediator manipulation] are randomly assigned in both time periods" (p. 153). They show that under this alternative design the average direct and indirect effects are separately identified when all variables are binary. This result, however, rests on the additional strong assumption that the causal effects of the treatment on both mediator and outcome as well as the causal effect of the mediator on the outcome are all monotonic. This monotonicity assumption is not made in our analysis below. A design identical to our crossover design was also discussed by Pearl (2001, p. 1574), albeit only in passing.

**Setup.** Let us denote the binary treatment variable in the first period by $T_i$. We write the potential mediator and outcome variables in the first period by $M_i(t)$ and $Y_{i1}(t', M_i(t))$, respectively. Then, the average indirect effect is given by $\bar{\delta}(t) = \mathbb{E}(Y_{i1}(t, M_i(1)) - Y_{i1}(t, M_i(0)))$ for $t = 0, 1$. During the second period of the experiment, the treatment status for each unit equals $1 - T_i$, and the value of the mediator, to which unit $i$ is assigned, equals the observed mediator value from the first period, $M_i$. Finally, the potential outcome in the second period can be written as $Y_{i2}(t, m)$ where the observed outcome is given by $Y_{i2} = Y_{i2}(1 - T_i, M_i)$. Since the treatment is randomized, the following assumption is automatically

satisfied under the crossover design.

ASSUMPTION 6 (RANDOMIZATION OF TREATMENT UNDER THE CROSSOVER DESIGN)

$$\{Y_{i1}(t,m), Y_{i2}(t',m), M_{i1}(t'') \; : \; t,t',t'' \in \{0,1\}, \; m \in \mathcal{M}\} \;\perp\!\!\!\perp\; T_i.$$

Like the parallel design, we make the consistency assumption. That is, the manipulation of the mediator in the second period does not directly affect the outcome, in the sense that the outcome variable would take the value that would naturally occur if the unit chose that particular value of the mediator without the manipulation. In addition to this consistency assumption, we also assume the absence of carryover effect as often done in the standard crossover trials. Specifically, we assume that the treatment administered in the first period does not affect the average outcome in the second period, as well as that there is no period effect (that is, the average potential outcomes remains the same in two periods). Formally, these key identifying assumptions can be stated as follows,

ASSUMPTION 7 (CONSISTENCY AND NO CARRYOVER EFFECTS UNDER THE CROSSOVER DESIGN)

$$\mathbb{E}(Y_{i1}(t, M_i(t))) \;=\; \mathbb{E}(Y_{i2}(t,m)) \quad \text{if} \quad M_i(t) = m,$$

for all $t = 0, 1$ and $m \in \mathcal{M}$.

This assumption allows us to write the expected values of potential outcomes in both periods simply as $\mathbb{E}(Y_i(t,m))$ for any $t$ and $m$. Notice that unlike the parallel design the consistency assumption only needs to hold in expectation, slightly relaxing Assumption 3. (If the assumption holds at the individual level, we can identify individual level direct and indirect effects.) Together, these assumptions allow researchers to observe two potential outcomes for each unit at different treatment conditions sequentially while holding the value of the mediator constant.

Assumption 7 might be violated if, for example, the exposure to the first treatment condition provides subjects with a reference point, which they then use in deciding how to respond to the subsequent treatment condition in the second experiment. Like Assumption 5, it is impossible to directly test Assumption 7; however, the assumption can be partially tested if we modify the second experiment to include an optional subgroup for each treatment group which do not receive any mediator manipulation. This test can be done by comparing the average observed outcome among each of these subgroups to the average outcome among the opposite treatment group in the first experiment. If the difference between these values are insignificant for both treatment conditions, the analyst can know that the no carryover effect (but not necessarily the consistency) assumption is plausible.

**Identification.** Under the crossover design, experimenters attempt to measure potential outcomes under different treatment and mediator values for each unit. This helps address the fundamental problem of identifying causal mechanisms discussed in Section 2. The next theorem summarizes the fact that under the crossover design the randomization of the treatment and the assumption of consistency and no carryover effects identify the average indirect effect.

THEOREM 2 (IDENTIFICATION UNDER THE CROSSOVER DESIGN) *Under Assumptions 6 and 7, the average indirect effect is identified and given by,*

$$\bar{\delta}(1) = \mathbb{E}(Y_{i1} \mid T_i = 1) - \mathbb{E}(Y_{i2} \mid T_i = 0),$$
$$\bar{\delta}(0) = \mathbb{E}(Y_{i2} \mid T_i = 1) - \mathbb{E}(Y_{i1} \mid T_i = 0).$$

Proof is straightforward, and therefore is omitted.

**Sharp bounds.** Under the crossover design, the assumption of consistency and no carryover effects is crucial. Without it, the sharp bounds on the average indirect effects would indeed be identical to those under the single experiment design given in equations (6) and (7) because the second experiment provides no relevant information. This is similar to the standard crossover design where the assumption of no carryover effect plays an essential role although the difference is that under the standard crossover design this assumption can be directly tested.

**Example.** In a landmark paper, Bertrand and Mullainathan (2004) conduct a randomized field experiment to test labor market discrimination against African Americans. The authors created fictitious resumes, some with typical White names and others with African American sounding names, thus only varying the perceived racial identity of applicants (the treatment $T_i$ which is equal to 1 if applicant $i$ is White and 0 if she is Black) while keeping their qualifications (the mediator $M_i$) constant. These resumes are then randomly sent to potential employers and callback rates for interviews are measured as the outcome variable of interest. The authors found that the resumes with White names are more likely to get callbacks than those with Black names.

Under the original experimental design, the researchers were able to estimate the average causal effect of manipulating applicants' race on callback rates, i.e., the average controlled direct effect $\bar{\eta}(m) = \mathbb{E}(Y_i(1, m) - Y_i(0, m))$ where $m$ represents the particular qualification specified in resumes. An alternative causal quantity of interest is the average direct effect of applicants' racial identity among African Americans, which represents the average increase in the callback rate if African American applicants were Whites but their qualifications stayed at the actual value, i.e., $\mathbb{E}(Y_i(1, M_i(0)) - Y_i(0, M_i(0)) \mid T_i = 0)$ (see

the discussion in Section 2.1). This quantity can thus be interpreted as the portion of the effect of race that does not go through the causal mechanism represented by qualifications.

The identification of this quantity is useful in order to isolate the degree to which African American job applicants are discriminated not based on qualifications but on their race. If the quantity is positive, then it may suggest the existence of racial discrimination in the labor market. The key difference between the two quantities is that the former is conditional on a particular qualification $m$ assigned by experimentalists while the latter holds applicants' qualifications constant at their actual observed values. The two quantities are different so long as the interaction between racial discrimination and the level of qualifications does exist, i.e., $\bar{\eta}(m) \neq \bar{\eta}(m')$ for $m \neq m'$. Indeed, Bertrand and Mullainathan (2004) found that the racial gap is larger when qualifications are higher, indicating that these two quantities are likely to diverge.

In this setting, the crossover design and its variants may be applicable. In the original study, the authors directly manipulated the qualifications by creating fictitious resumes (i.e., setting $M_i$ to some arbitrary $m$). Instead, we could sample actual resumes of African American job applicants to obtain $M_i(0)$. Sending these resumes without any modification will allow us to identify $\mathbb{E}(Y_i(0, M_i(0)) \mid T_i = 0)$. We could then change the names of applicants to White sounding names in order to identify the counterfactual outcome $\mathbb{E}(Y_i(1, M_i(0)) \mid T_i = 0)$ without changing the other parts of the resumes (i.e., holding $M_i$ constant at $M_i(0)$). The consistency assumption is plausible here so long as potential employers are kept unaware of the name manipulation as done in the original study. The no carryover effect assumption may be problematic if the same resume with different names is sent to the same potential employer over two time periods. Fortunately, this problem can be overcome by sending these resumes to different (randomly matched) employers at the same time, thereby averaging over the distribution of potential employers. This strategy is effective because the assumption of consistency and no carryover effects only need to hold in expectation. Thus, researchers will be able to infer how much of labor market discrimination can be attributable to race rather than qualification of a job applicant.

## 4    Experimental Designs with Imperfect Manipulation

Although the above two experimental designs yield greater identification power than the standard single experiment design, the direct manipulation of the mediator is often difficult in practice. Moreover, even when such manipulations are possible, the consistency assumptions may not be credible especially if an explicit intervention must be given to control the value of the mediator. To address this issue, we consider new experimental designs that generalize the previous two designs by allowing for the imperfect

manipulation of the mediator. These designs might be useful in the situations where researchers can only encourage (rather than assign) experimental subjects to take a particular value of the mediator. Such randomized encouragement has been previously studied in the context of identifying treatment effects (Angrist *et al.*, 1996) and principal strata direct effects (Mattei and Mealli, 2011).

Here, we consider the use of randomized encouragement for the identification of causal mechanisms, which may be preferable even when the direct manipulation is possible because subtle encouragement tends to increase the credibility of the consistency assumption about the mediator manipulation. Note that our use of encouragement differs from some previous works in the literature where the treatment variable is used as an instrumental variable for the mediator under the standard design with the assumption of no direct effect of the treatment on the outcome (e.g., Jo, 2008; Sobel, 2008). In contrast, we allow for the direct effect of the treatment on the outcome, the identification of which is typically a primary goal of causal mediation analysis.

## 4.1 The Parallel Encouragement Design

The *parallel encouragement design* is a generalization of the parallel design where the manipulation of the mediator can be imperfect. Thus, instead of directly manipulating the mediator in the second experiment, we randomly encourage subjects to take a particular value of the mediator.

**Setup.** Formally, let $Z_i$ represent the ternary encouragement variable where it is equal to $1$ ($-1$) if subject $i$ is positively (negatively) encouraged and is equal to $0$ if no such encouragement is given. Then, the potential value of the mediator can be written as the function of both the treatment and encouragement, i.e., $M_i(t, z)$ for $t = 0, 1$ and $z = -1, 0, 1$. Similarly, the potential outcome is a function of the encouragement as well as the treatment and the mediator, i.e., $Y_i(t, m, z)$. Then, the observed values of the mediator and the outcome are given by $M_i(T_i, Z_i)$ and $Y_i(T_i, M_i(T_i, Z_i), Z_i)$, respectively. For the sake of simplicity, we assume that the mediator is binary. The randomization of the treatment and the encouragement implies that the following independence assumption holds,

ASSUMPTION 8 (RANDOMIZATION OF THE TREATMENT AND THE ENCOURAGEMENT) *For $m = 0, 1$,*

$$\{Y_i(t, m, z), M_i(t', z') \ : \ t, t' \in \{0, 1\}, \ z, z' \in \{-1, 0, 1\}\} \quad \perp\!\!\!\perp \quad \{T_i, Z_i\}.$$

Here, both the treatment and encouragement are assumed to be under perfect control of the analyst and thus conditioning on pre-treatment or pre-encouragement covariates is not required.

Furthermore, as done in the standard encouragement design, we make two assumptions (the "exclusion restriction" and "monotonicity"; see Angrist *et al.*, 1996). First, we assume the encouragement affects

the outcome only through the mediator. This represents the consistency assumption under the parallel encouragement design. Second, we assume that the encouragement monotonically affects the mediator. That is, there exist no "defiers" who behave exactly opposite to the encouragement. Without loss of generality, these two assumptions can be formalized as,

ASSUMPTION 9 (CONSISTENCY UNDER THE PARALLEL ENCOURAGEMENT DESIGN) *For all* $t = 0, 1$ *and* $z, z' = -1, 0, 1$.

$$Y_i(t, M_i(t, z), z) = Y_i(t, M_i(t, z'), z') \quad \text{if} \quad M_i(t, z) = M_i(t, z').$$

ASSUMPTION 10 (MONOTONICITY) *For* $t = 0, 1$,

$$M_i(t, 1) \geq M_i(t, 0) \geq M_i(t, -1).$$

Because the potential outcomes do not directly depend on the value of the encouragement under Assumption 9, we can write them simply as $Y_i(t, m)$ for any $t$ and $m$.

Under the parallel encouragement design, our quantity of interest is the average indirect effects for "compliers" which refer to those who are affected by either the positive or negative encouragement in the intended direction under a given treatment status. We note that compliance status may depend on how the encouragement is implemented, as in other settings. The quantity we focus on is analogous to the average complier causal effects, which can be identified under the standard encouragement design (Angrist *et al.*, 1996). We can formally define the average complier indirect effects under this setting as follows,

$$\bar{\delta}^*(t) = \mathbb{E}(Y_i(t, M_i(t, 0)) - Y_i(t, M_i(t', 0)) \mid (M_i(t, -1), M_i(t, 0), M_i(t, 1)) \in \{(0, 0, 1), (0, 1, 1)\}),$$

for $t = 0, 1$ and $t \neq t'$.

**Sharp bounds.** Given this setup, we study the identification power of the parallel encouragement design again using a bounds approach. Again, for the sake of simplicity and comparison with the other designs, we focus on the situation where the outcome is also binary. In this case, under Assumptions 8, 9, and 10, the sharp bounds can be derived numerically using a standard linear programming routine. Appendix A.5 provides the details of the derivation of the sharp bounds on the average complier indirect effects.

**Example.** As a potential application of the parallel encouragement design, we consider the media framing experiment by Brader *et al.* (2008) which used the single experiment design. As discussed in Section 2.3, the mediator of interest in this study is the level of anxiety, a psychological factor that is difficult to manipulate directly. While this prevents researchers from using the parallel design, the parallel encouragement design may be applicable to this type of psychological experiments. Under the parallel encouragement design, we first randomly split the sample into two groups. Then, for one group, the treatment is

randomly assigned but no manipulation of mediator is conducted. For the other group, experimenters randomize the treatment and the indirect manipulation to change the level of anxiety. Since the manipulation of a psychological factor is likely to be imperfect, this constitutes the parallel encouragement design.

In the psychological literature, there exist several ways to indirectly manipulate emotion. A common method is the autobiographical emotional memory task, where participants write about an event in their life that made them feel a particular emotion (e.g., Lerner and Keltner, 2001). Using such a task to manipulate anxiety would satisfy the consistency assumption (Assumption 9) if, for any given treatment assignment and anxiety level, a subject reports the same immigration preference regardless of whether their anxiety level was manipulated or chosen by the subject. The assumption is violated if, for example, a subject interprets the task of writing a negative experience as an indication that the experiment is concerned about negative aspects of immigration. Protocol to minimize such problems (e.g., by not mentioning immigration or other ethnicity in task instructions) can help make the consistency assumption more plausible.

The other key assumption of monotonicity (Assumption 10) will be violated if there are any subjects whose anxiety level would be decreased by the writing task that was purported to increase anxiety. This could be a serious concern because it has been found that the act of expressing a certain emotion can have a cathartic effect on the emotion and decrease its intensity in one's mind. Careful choice of a writing task will thus be a crucial factor in successfully implementing this design in practice.

## 4.2   The Crossover Encouragement Design

It is also possible to generalize the crossover design described in Section 3.2 so that the imperfect manipulation of the mediator is allowed. Under this *crossover encouragement design*, after the treatment is randomized, the value of the mediator and then optionally the value of the outcome are observed for each unit in both treatment and control groups. Thus, the first period remains unchanged from the crossover design except that the measurement of the outcome variable is no longer required for identification (though it is recommended as discussed below). The second period, however, is different. After assigning each unit to the treatment condition opposite to their first period status, the experimenter encourages randomly selected units so that their mediator equals its observed value from the first period.

As shown below, under some assumptions this design identifies average indirect effects for the specific subpopulation we call the *pliable* units. While the information from the first period is primarily used to determine the direction of encouragement given in the second period, the (randomly selected) group that receives no encouragement in the second period is used to learn about the proportion of these pliable units,

or those who would change behavior in response to the encouragement. We then combine this with other information obtained from the second period to identify causal mechanisms among the pliables.

**Setup.** Formally, let $V_i$ represent the randomized binary encouragement variable where $V_i = 1$ indicates that unit $i$ receives the encouragement to take the same value of the mediator during the second period as the first period. $V_i = 0$ represents the absence of such encouragement (i.e., do nothing). Then, the potential values of the mediator during the second period can be written as $M_{i2}(t, v)$ under the treatment status $t$ and the encouragement status $v$ of this period. Similarly, we write the potential outcomes for the second period as $Y_{i2}(t, m, v)$ where $t$ and $m$ represent the values of the treatment and the mediator during the second period, and $v$ denotes the encouragement status. As before, we assume consistency in that the indirect manipulation of the mediator through the encouragement has no direct effect on the outcome other than through the resulting value of the mediator. This assumption, together with the assumption of no carryover effect (for both the mediator and the outcome), can be formalized as,

ASSUMPTION 11 (CONSISTENCY AND NO CARRYOVER EFFECTS UNDER CROSSOVER ENCOURAGE-MENT DESIGN) *For all $t, t', v = 0, 1$,*

$$M_{i1}(t) = M_{i2}(t, 0) \quad and \quad Y_{i1}(t, M_{1i}(t')) = Y_{i2}(t, M_{i2}(t, v), v) \quad if \quad M_{i1}(t') = M_{i2}(t, v).$$

The first part of this assumption allows both the potential mediator in the first period as well as the second-period mediator when $V_i = 0$ to be written simply as $M_i(t)$ for any $t$. Similarly, the notation for the potential outcomes in both periods can be simplified to $Y_i(t, m)$.

One advantage of the crossover encouragement design is that unlike the crossover design, researchers can test observable implications of the consistency and no carryover effects assumptions. First, it is possible to test whether the first equality in Assumption 11 holds on average by comparing $\mathbb{E}(M_{i1} \mid T_i = t)$ with $\mathbb{E}(M_{i2} \mid T_i = 1-t, V_i = 0)$ for $t = 0, 1$. This is because these two quantities are equal to the expected values of the two potential mediator values in the first equality in Assumption 11 when both the treatment and encouragement are randomized. Second, the second equality in Assumption 11 can be partially tested by comparing $\mathbb{E}(Y_{i1} \mid T_i = t, M_{i1} = m)$ with $\mathbb{E}(Y_{i2} \mid T_i = 1 - t, M_{i2} = m, V_i = 0)$ for $m = 0, 1$. This is because these two quantities are equal to $\mathbb{E}(Y_{i1}(t, m) \mid M_{i1}(t) = m)$ and $\mathbb{E}(Y_{i2}(t, m, 0) \mid M_{i2}(t, 0) = m)$, respectively, and thus under the assumption that the first equality in Assumption 11 is true the comparison yields a test whether the second equality holds in expectation when $v = 0$. However, it should be noted that this procedure has no implication for the case in which $v = 1$ and thus cannot be used for testing whether there is a direct effect of encouragement itself on the outcome. Nevertheless, we recommend

21

measuring the first period outcome because it allows testing whether there is any carryover effect and it often involves little additional cost.

In addition to these assumptions, which are essentially equivalent to the assumptions made under the crossover design, we rely upon the following monotonicity assumption as done under the parallel encouragement design. In particular, we assume that no unit would take the value of the mediator equal to its observed value from the first period *only when* they are *not* encouraged. When the mediator is binary, the assumption can be written formally as,

ASSUMPTION 12 (NO DEFIER) *For any $t = 0, 1$ and $m \in \mathcal{M}$,*

$$\Pr(M_{i2}(1 - t, 0) = m, M_{i2}(1 - t, 1) = 1 - m \mid M_{i1} = m, T_i = t) = 0.$$

Finally, the randomization of the treatment and the encouragement implies the following,

ASSUMPTION 13 (RANDOMIZATION OF TREATMENT AND ENCOURAGEMENT) *For any $m^* \in \mathcal{M}$ and $t^* \in \{0, 1\}$,*

$$\{Y_{i1}(t, m), Y_{i2}(t', m, v), M_{i1}(t_1), M_{i2}(t_2) \ : \ t, t', t_1, t_2, v \in \{0, 1\}, \ m \in \mathcal{M}\} \perp\!\!\!\perp T_i$$
$$\{Y_{i2}(t', m, v), M_{i2}(t_2) \ : \ t', t_2, v \in \{0, 1\}, \ m \in \mathcal{M}\} \perp\!\!\!\perp V_i \mid M_{i1} = m^*, T_i = t^*.$$

**Identification.** Under these assumptions and binary mediator and outcome variables, one can identify the average indirect effect but only for a subset of the population who can be successfully manipulated by the experimenter via the encouragement. These *pliable* units are those for whom the value of the mediator in the second experiment is the same as the value in the first experiment *only if* they are encouraged. We focus on this subpopulation because as in instrumental variable methods, this design is not informative about those who are not affected by the encouragement. Formally, the average indirect effects among pliable units are defined as,

$$\bar{\delta}_P(t) \ \equiv \ \mathbb{E}(Y_i(t, M_i(1)) - Y_i(t, M_i(0)) \mid M_{i2}(t, 0) = 1 - M_{i1}(1 - t), M_{i2}(t, 1) = M_{i1}(1 - t)),$$

for $t = 0, 1$. In Appendix A.6, we prove that these quantities are identified under the crossover encouragement design with Assumptions 11–13.

**Example.** As a potential application of the crossover encouragement design, we consider the recent survey experiment by Hainmueller and Hiscox (2010) about the effects of issue framing on preferences towards immigration. The authors study how immigration preferences of low income U.S. citizens are influenced by whether they are asked about high or low skill immigrants. One of the hypotheses they consider is that competition over public resources between natives and immigrants leads to greater opposition towards immigration. If this is true, natives will form negative expectations about the impact of

immigrants on access to public services. While Hainmueller and Hiscox (2010) are unable to directly test this mechanism, a modification of their original experimental design may permit this.

The study used the standard $2 \times 2$ crossover design where survey respondents were first randomly asked to consider either high or low skill immigrants and then express their policy preferences about increasing immigration. Two weeks later, the same respondents were surveyed again, except that they were asked about the other skill group, thereby reversing the treatment. The authors found that expressed preferences about immigration differ substantially depending on whether respondents were asked about low or high skill immigrants. Low income respondents who opposed immigration after being exposed to the low skill immigrant frame tended to become favorable when asked to consider high skill immigrants.

To investigate the hypothesized causal mechanism, the original experimental design may be modified as follows. Following the framing about high ($T_i = 1$) or low skill immigrants ($T_i = 0$), we would ask respondents for their expectations about the ease of access to public services or the availability of welfare services in the future ($M_{i1}$). In the second experiment, for the same respondents, the skill treatment would be reversed but the experiment would include an additional manipulation designed to change expectations about public service access in the same direction as was observed in the first experiment ($V_i$). For example, if someone in the first experiment received the low skill frame and stated that they expect future access to public services to decline, then the second period manipulation of these expectations could be in the form of a news story reporting that state budgets were unlikely to be able to support future public service spending. Following this manipulation of the mediating variable the respondents would be asked again for their expectations about public service access ($M_{i2}$) and the preferences over immigration flows ($Y_{i2}$).

Is the no carryover effect assumption likely to be met in this example? In the original experiment Hainmueller and Hiscox (2010) staggered the two waves of their survey by approximately two weeks and found little carryover effects. The long washout period in their design makes the no carryover assumption more plausible. As for the consistency assumption, the key question is whether the use of a news story has a direct influence on subjects' preferences over immigration other than through the hypothesized mechanism. The answer to this question perhaps requires additional investigation.

# 5  A Numerical Example

We now illustrate some of our analytical results using a numerical example based on the media framing experiment by Brader *et al.* (2008). As described earlier, the substantive question of interest is whether the effect of media framing on subjects' immigration preference is mediated by changes in the level of

|                         | Treatment |       | Control |       |             |
|-------------------------|-----------|-------|---------|-------|-------------|
| Response Variables      | Mean      | S.D.  | Mean    | S.D.  | ATE (s.e.)  |
| Anxiety Level           | 0.603     | 0.493 | 0.328   | 0.471 | 0.275 (0.069) |
| Opposition to Immigration | 0.824   | 0.384 | 0.641   | 0.481 | 0.182 (0.058) |
| Sample Size             | 68        |       | 198     |       |             |

Table 1: Descriptive Statistics and Estimated Average Treatment Effects from the Immigration Experiment. The middle four columns show the mean and standard deviation of the mediator and outcome variables for each group. The last column reports the estimated average causal effects of the treatment (Latino image and negative tone) as opposed to the control condition on the hypothesized mediator and outcome variables along with their standard errors. The estimates suggest that the treatment affected each of these variables in the expected directions.

anxiety. Table 1 reports descriptive statistics and estimated average treatment effects computed from the original experimental results. Respondents in the treatment condition (Latino image and negative tone) exhibited significantly higher levels of anxiety and opposition to immigration than did respondents in the other conditions, leading to the estimated average treatment effects significantly greater than zero.

Here we conduct a simulation study using these results as a starting point. We first generate a population distribution of the potential outcomes and mediators as well as the compliance types with respect to the encouragement. To ensure the comparability of our simulated data with the distribution of observed variables, we randomly draw the joint probabilities of these causal types from a prior distribution which is consistent with the original data. The resulting population distribution is thus generated in such a way that the observed data in Table 1 could have come from this data generating process. We then randomly assign both the experimental condition for the parallel design ($D_i$) and the encouragement status ($Z_i$) to this simulated population. The resulting proportions of compliers (as defined in Section 4.1) are $0.730$ for the treatment group and $0.392$ for the control group. Finally, the observed values of the mediator and outcome under these designs are determined based on these two variables.

Figure 2 presents the sharp bounds on the average indirect effects for $t = 1$ (left panel) and $t = 0$ (right panel) under different experimental designs calculated from the simulated population. In both panels, the top three solid circles represent the true values of the average indirect effects ($\bar{\delta}(1) = 0.301$ and $\bar{\delta}(0) = -0.045$) and the bottom circles indicate the complier average indirect effects ($\bar{\delta}^*(1) = 0.392$ and $\bar{\delta}^*(0) = 0.014$). The horizontal bars represent the bounds under (from top to bottom) the single experiment design, parallel design, and parallel encouragement design. For the parallel encouragement design, we present the sharp bounds for both $\bar{\delta}(t)$ and $\bar{\delta}^*(t)$. The graphs illustrate the relative identification powers of these experimental designs. Under the single experiment design, the sharp bounds are wide for
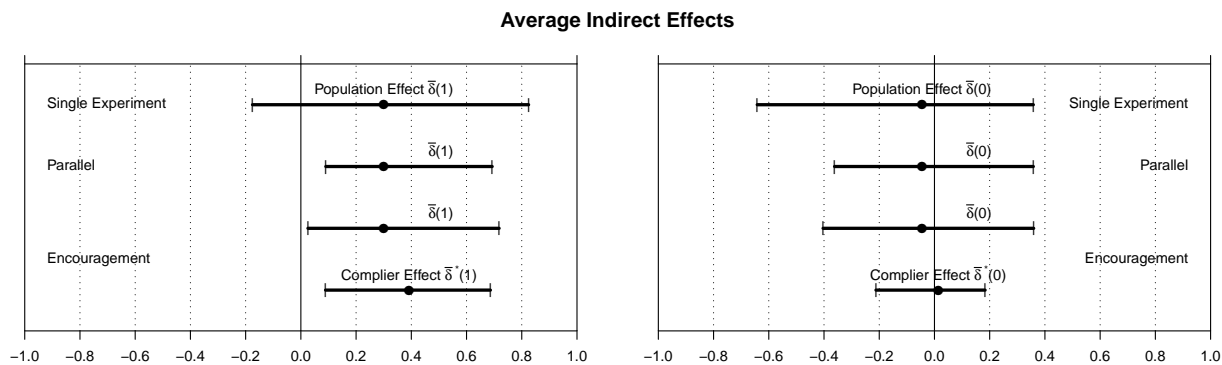
Figure 2: Identification Power of Alternative Experimental Designs. The graphs show the sharp bounds on the average indirect effects for the treatment (left panel) and control (right panel) conditions calculated on the basis of the hypothetical population distribution we generated. In each panel, the solid circles represent the true values of $\bar{\delta}(t)$ (top three) and $\bar{\delta}^*(t)$ (bottom). The horizontal bars indicate the sharp bounds under (from top to bottom) the single experiment design, parallel design, and parallel encouragement design. The graphs show improved identification powers of the new designs compared to the traditional single experiment design.

both $\bar{\delta}(1)$ and $\bar{\delta}(0)$ and include zero ($[-0.175, 0.825]$ and $[-0.642, 0.359]$, respectively). In contrast, the parallel design identifies the sign of $\bar{\delta}(1)$ to be positive without relying on any untestable assumption ($[0.090, 0.693]$), although it unsurprisingly fails to identify the sign of $\bar{\delta}(0)$ ($[-0.362, 0.358]$), whose true value is close to zero.

The parallel encouragement design is slightly less informative about the average indirect effects than the parallel design but nonetheless identifies the sign of $\bar{\delta}(1)$, with the sharp bounds of $[0.026, 0.718]$ and $[-0.403, 0.359]$ for $\bar{\delta}(1)$ and $\bar{\delta}(0)$, respectively. Moreover, the parallel encouragement design is even more informative about the complier average indirect effects; the sharp bounds for $\bar{\delta}^*(t)$ are narrower than any of the bounds for the average indirect effects for both $t = 1$ and $t = 0$ and do not include zero for the former ($[0.089, 0.686]$ and $[-0.212, 0.183]$). In sum, for our simulated population based on the experimental data of Brader *et al.* (2008), the parallel design and parallel encouragement design are substantially more informative about the average indirect effects than the standard single experiment design.

# 6 Concluding Remarks

The identification of causal mechanisms is at the heart of scientific research. Applied researchers in a variety of scientific disciplines seek to explain causal processes as well as estimating causal effects. As a consequence, experimental research has been often criticized as a black-box approach that ignores causal mechanisms. Despite this situation, both methodologists and experimentalists have paid relatively little

attention to an important question of how to design an experiment in order to empirically test the existence of hypothesized causal mechanisms. In this paper, we answer this question by proposing alternative experimental designs and analyzing the identification power of each design under various assumptions.

In applied research, the most dominant approach has been the *single experiment design* where only the treatment variable is randomized. The fundamental difficulty of this approach is that like in observational studies the absence of unobserved confounders is required for identification but this is never guaranteed to hold in practice. To overcome this limitation, we propose several alternative experimental designs that involve some kind of manipulation of mediator. Some designs we consider require the direct manipulation of mediator while others allow for the indirect and imperfect manipulation.

The key assumption under these experimental designs is that the action of manipulating the mediator does not directly affect the outcome (other than through the fact that the mediator takes a particular value). To satisfy this consistency assumption, the mediator must be manipulated in a way that experimental units behave as if they chose the mediator value on their own. This may appear to suggest that any experimental design involving some kind of manipulation of the mediator is potentially of limited use for the analysis of causal mechanisms. However, through the discussion of recent social science experiments, we have shown that such manipulation may become possible through technological advances in experimental methodology (e.g., the neuroscience experiment discussed in Section 3.1) as well as the creativity on the part of experimenters (e.g., the labor market discrimination experiment discussed in Section 3.2).

The proposed methodology emphasizes the identification assumptions that are directly linked to experimental design rather than those on the characteristics of experimental units. While experimenters can only play a passive role when making the second type of assumptions, they can improve the validity of the first type of assumptions through careful design and implementation of experiments. Thus, we hope that experimental designs considered in this paper will open up the possibilities to experimentally identify causal mechanisms through clever manipulations and future technological developments. While in this paper we draw only on social science examples, we believe that our designs could be used with slight or no modification for other settings, such as large-scale medical trials or public policy evaluations.

# A Appendix

## A.1 Relation to Geneletti (2007)'s Indirect Effects

Based on a non-counterfactual framework of causal inference, Geneletti shows how to identify alternative quantities called the "generated direct effect" and "indirect effect," which together add up to the average

causal effect $\bar{\tau}$. The relative advantages and disadvantages of the counterfactual versus non-counterfactual approaches to causal inference are beyond the scope of the current paper (see Dawid, 2000). However, it appears that Geneletti's indirect effect can be reexpressed using potential outcomes in the following way, $\bar{\delta}^{\dagger}(t) = \mathbb{E}(Y_i(t, M_1) - Y_i(t, M_0) \mid F_{M_0} = F_{M_i(0)}, F_{M_1} = F_{M_i(1)})$, for $t = 0, 1$ where $F_X$ represents the distribution of random variable $X$. This differs from the average natural indirect effect $\bar{\delta}$, which for the sake of comparison can be rewritten as $\mathbb{E}(Y_i(t, M_1) - Y_i(t, M_0) \mid M_0 = M_i(0), M_1 = M_i(1))$.

The difference between $\bar{\delta}^{\dagger}(t)$ and $\bar{\delta}(t)$ is rather subtle but important. For illustration, we use Geneletti's example (see Section 3.1.2(b)) about a drug treatment for a particular disease that may trigger headaches as side effect. In the example, aspirin is taken by patients to alleviate the headaches and acts as a mediator for the outcome of disease prognosis. In this context, the natural indirect effect represents the causal effect of the drug on the disease prognosis that is transmitted through changes in patients' aspirin intake following their administration of the treatment drug. In contrast, Geneletti's indirect effect represents the causal effect of a hypothetical intervention where aspirin intake is randomly assigned according to the population distribution of natural levels of aspirin under the treatment and control conditions. Therefore, this alternative quantity does not directly correspond to a causal process unless units in the population are assumed to be exchangeable (a difficult assumption to maintain given the heterogeneity of human population). Our approach, on the other hand, avoids this exchangeability assumption and develops experimental designs that help identify causal mechanisms under less stringent assumptions.

## A.2 Sharp Bounds under the Single Experiment Design

We present the sharp bounds on the average indirect effects under the single experiment design. These bounds can be obtained by solving a linear optimization problem with respect to $\bar{\delta}(1)$ and $\bar{\delta}(0)$ under the constraints implied by Assumption 1 alone. Here we take a simpler alternative approach which uses the equality given in Section 2.1, $\bar{\tau} = \bar{\delta}(t) + \bar{\zeta}(1-t)$. That is, we subtract the sharp bounds on $\bar{\zeta}(1-t)$ derived by Sjölander (2009) from the average total effect, which is identified under Assumption 1, to obtain the

following bounds on $\bar{\delta}(t)$ for $t = 0, 1$,

$$\max \left\{ \begin{array}{c} -P_{001} - P_{011} \\ -P_{011} - P_{010} - P_{110} \\ -P_{000} - P_{001} - P_{100} \end{array} \right\} \leq \bar{\delta}(1) \leq \min \left\{ \begin{array}{c} P_{101} + P_{111} \\ P_{010} + P_{110} + P_{111} \\ P_{000} + P_{100} + P_{101} \end{array} \right\}, \quad (6)$$

$$\max \left\{ \begin{array}{c} -P_{100} - P_{110} \\ -P_{011} - P_{111} - P_{110} \\ -P_{001} - P_{101} - P_{100} \end{array} \right\} \leq \bar{\delta}(0) \leq \min \left\{ \begin{array}{c} P_{000} + P_{010} \\ P_{011} + P_{111} + P_{010} \\ P_{000} + P_{001} + P_{101} \end{array} \right\}, \quad (7)$$

where $P_{ymt} = \Pr(Y_i = y, M_i = m \mid T_i = t, D_i = 0)$.

## A.3    Proof of Theorem 1

We begin by noting that both $\mathbb{E}(Y_i(t, M_i(t)))$ and $\mathbb{E}(Y_i(t, m))$ are identified for any $t$ and $m$ under Assumptions 1, 3 and 4. The former can be identified from the first experiment by $\int \mathbb{E}(Y_i \mid T_i = t, X_i = x, D_i = 0) dF_{X_i \mid D_i = 0}(x)$ and the latter from the second experiment by $\int \mathbb{E}(Y_i \mid T_i = t, M_i = m, X_i = x, D_i = 1) dF_{X_i \mid D_i = 1}(x)$. Thus, by following the proof of Theorem 2.1 of Robins (2003), under Assumption 5 the average indirect effect is identified and given by $\bar{\delta}(1) = \bar{\delta}(0) = \bar{\tau} - \bar{\zeta}(t)$ where $\bar{\zeta}(t) = \mathbb{E}(Y_i(1, m) - Y_i(0, m))$ for any $m \in \mathcal{M}$. $\qquad \square$

## A.4    Sharp Bounds under the Parallel Design

We derive the large-sample sharp bounds on the average indirect effects under the parallel design with binary mediator and outcome variables. For $\bar{\delta}(1)$, we just need to derive the sharp bounds on $\mathbb{E}\{Y_i(1, M_i(0))\}$ because $\mathbb{E}\{Y_i(0, M_i(0))\}$ is identified as $\Pr(Y_i = 1 \mid T_i = 0, D_i = 0)$. From equation (5), the former quantity can be decomposed as,

$$\mathbb{E}\{Y_i(1, M_i(0))\} = \sum_{y=0}^{1} \sum_{m=0}^{1} (\pi_{1ym1} + \pi_{y1m1})$$

where $\pi_{y_1 y_0 m_1 m_0} = \Pr(Y_i(1, 1) = y_1, Y_i(1, 0) = y_0, M_i(1) = m_1, M_i(0) = m_0) \geq 0$ with the constraint $\sum_{y_1=0}^{1} \sum_{y_0=0}^{1} \sum_{m_1=0}^{1} \sum_{m_0=0}^{1} \pi_{y_1 y_0 m_1 m_0} = 1$. This quantity can be maximized or minimized via standard linear programming techniques. Thus, we can derive the sharp bounds by finding the optima of this

quantity under the following constraints implied by the experimental design,

$$\Pr(M_i = 1 \mid T_i = 0, D_i = 0) = \sum_{y_1=0}^{1}\sum_{y_0=0}^{1}\sum_{m=0}^{1} \pi_{y_1 y_0 m 1}$$

$$\Pr(M_i = 1 \mid T_i = 1, D_i = 0) = \sum_{y_1=0}^{1}\sum_{y_0=0}^{1}\sum_{m=0}^{1} \pi_{y_1 y_0 1 m}$$

$$\Pr(Y_i = 1, M_i = m \mid T_i = 1, D_i = 0) = \begin{cases} \sum_{y_0=0}^{1}\sum_{m_0=0}^{1} \pi_{1 y_0 m m_0} & \text{if } m = 1 \\ \sum_{y_1=0}^{1}\sum_{m_0=0}^{1} \pi_{y_1 1 m m_0} & \text{if } m = 0 \end{cases}$$

$$\Pr(Y_i = 1 \mid M_i = m, T_i = 1, D_i = 1) = \begin{cases} \sum_{y_0=0}^{1}\sum_{m_1=0}^{1}\sum_{m_0=0}^{1} \pi_{1 y_0 m_1 m_0} & \text{if } m = 1 \\ \sum_{y_1=0}^{1}\sum_{m_1=0}^{1}\sum_{m_0=0}^{1} \pi_{y_1 1 m_1 m_0} & \text{if } m = 0 \end{cases}$$

The sharp bounds on $\bar{\delta}(1)$ can then be obtained by combining the above with the already identified quantity, $\mathbb{E}\{Y_i(0, M_i(0))\}$. A similar calculation yields the sharp bounds on $\bar{\delta}(0)$.

The resulting sharp bounds under Assumptions 1, 3 and 4 are given by,

$$\max \begin{cases} -P_{001} - P_{011} \\ -P_{011} - P_{010} - P_{110} - P_{001} + Q_{001} \\ -P_{000} - P_{001} - P_{100} - P_{011} + Q_{011} \\ -P_{001} - P_{011} + Q_{001} - Q_{111} \\ -P_{001} + P_{101} - Q_{101} \\ -P_{011} + P_{111} - Q_{111} \end{cases} \leq \bar{\delta}(1) \leq \min \begin{cases} P_{101} + P_{111} \\ P_{010} + P_{110} + P_{101} + P_{111} - Q_{101} \\ P_{000} + P_{100} + P_{101} + P_{111} - Q_{111} \\ P_{101} + P_{111} + Q_{001} - Q_{111} \\ P_{111} - P_{011} + Q_{011} \\ P_{101} - P_{001} + Q_{001} \end{cases}, \quad (8)$$

$$\max \begin{cases} -P_{100} - P_{110} \\ -P_{011} - P_{111} - P_{110} - P_{100} + Q_{000} \\ -P_{001} - P_{101} - P_{100} - P_{110} + Q_{110} \\ -P_{100} - P_{110} + Q_{100} - Q_{010} \\ -P_{110} + P_{010} - Q_{010} \\ -P_{100} + P_{000} - Q_{100} \end{cases} \leq \bar{\delta}(0) \leq \min \begin{cases} P_{000} + P_{010} \\ P_{011} + P_{111} + P_{010} + P_{000} - Q_{100} \\ P_{000} + P_{001} + P_{101} + P_{010} - Q_{010} \\ P_{000} + P_{010} + Q_{100} - Q_{010} \\ P_{010} - P_{110} + Q_{110} \\ P_{000} - P_{100} + Q_{100} \end{cases}, \quad (9)$$

where $P_{ymt} \equiv \Pr(Y_i = y, M_i = m \mid T_i = t, D_i = 0)$, $Q_{ymt} \equiv \Pr(Y_i = y \mid M_i = m, T_i = t, D_i = 1)$.

As expected, these bounds are at least as informative as the bounds under the single experiment design. This can be shown formally by deriving the sharp bounds on $Q_{ymt}$ under the single experiment design and then substituting them into equations (8) and (9). For example, under the single experiment design, we have $P_{001} \leq Q_{001} \leq P_{001} + P_{011} + P_{111}$, $P_{011} \leq Q_{011} \leq P_{001} + P_{101} + P_{011}$, and $-P_{101} - P_{111} \leq Q_{001} - Q_{111} \leq P_{001} + P_{011}$. Thus, under this design, the expression for the lower bound of $\bar{\delta}(1)$ given in equation (8) reduces to that of equation (6).

Moreover, the above expression of the bounds imply that unlike the single experiment design, the parallel design can sometimes identify the sign of the average indirect effects. However, there exists a tradeoff between the informativeness of the lower bound and that of the upper bound. For example, if the values of $Q_{001}$ and $Q_{011}$ are large (small), then the lower bound of $\bar{\delta}(1)$ will be large (small) but so will be the upper bound of $\bar{\delta}(1)$.

## A.5 Sharp Bounds under the Parallel Encouragement Design

The average complier indirect effect can be decomposed and expressed with respect to the principal strata probabilities as,

$$\delta^*(t) = \frac{\sum_{m'_{-1}} \sum_{m'_1} \left( \psi^{01}_{m'_{-1}0m'_1011} + \psi^{10}_{m'_{-1}1m'_1001} - \psi^{01}_{m'_{-1}1m'_1001} - \psi^{10}_{m'_{-1}0m'_1011} \right)}{\Pr(M_i(t,-1)=0, M_i(t,0)=M_i(t,1)=1) + \Pr(M_i(t,-1)=M_i(t,0)=0, M_i(t,1)=1)},$$

(10)

where $\psi^{y_0y_1}_{m'_{-1}m'_0m'_1m_{-1}m_0m_1} = \Pr(Y_i(t,0)=y_0, Y_i(t,1)=y_1, M_i(t',-1)=m'_{-1}, M_i(t',0)=m'_0, M_i(t',1)=m'_1, M_i(t,-1)=m_{-1}, M_i(t,0)=m_0, M_i(t,1)=m_1) \geq 0$, with the constraint

$$\sum_{y_0=0}^{1} \sum_{y_1=0}^{1} \sum_{m'_{-1}=0}^{1} \sum_{m'_0=0}^{1} \sum_{m'_1=0}^{1} \sum_{m_{-1}=0}^{1} \sum_{m_0=0}^{1} \sum_{m_1=0}^{1} \psi^{y_0y_1}_{m'_{-1}m'_0m'_1m_{-1}m_0m_1} = 1.$$

Note that the denominator can be identified from the observed data and expressed as $P^{\dagger}_{00t} + P^{\dagger}_{10t} - P^*_{00t} - P^*_{10t}$, where $P^{\dagger}_{ymt} = \Pr(Y_i=y, M_i=m \mid T_i=t, Z_i=-1)$. Thus, the sharp bounds on $\delta^*(t)$ can be obtained by maximizing and minimizing the numerator via a standard linear programming algorithm subject to the following linear constraints implied by the experimental design,

$$\Pr(Y_i=y, M_i=m \mid T_i=t, Z_i=z) = \Pr(Y_i(t,m)=y, M_i(t,z)=m),$$
$$\Pr(M_i=1 \mid T_i=t', Z_i=z) = \Pr(M_i(t',z)=1),$$

for $y=0,1$, $m=0,1$ and $z=0,1$.

Depending on the context of one's research, it may be possible to make additional assumptions about causal relationships between the treatment, mediator and outcome. Such assumptions can be incorporated as long as they are expressed as linear functions of the principal strata. For example, one may want to make the no interaction effect assumption (Assumption 5). This assumption can be written as $\Pr(Y_i(t,1) - Y_i(t,0) \neq Y_i(t',1) - Y_i(t',0)) = 0$ and, since this is linear in principal strata, the sharp bounds on the average indirect effects with this additional assumption can be derived using the above framework. $\square$

## A.6 Identification under the Crossover Encouragement Design

We prove the following identification result under the crossover encouragement design.

THEOREM 3 (IDENTIFICATION UNDER THE CROSSOVER ENCOURAGEMENT DESIGN) *Under Assumptions 11–13, the average indirect effect among the pliable units is identified and given by,*

$$
\begin{aligned}
\bar{\delta}_P(1-t) &= (1-2t)\left[\{\Lambda_{t00}\Gamma_{t010} + (1-\Lambda_{t00})\Gamma_{t000} - \Lambda_{t01}\Gamma_{t011} - (1-\Lambda_{t01})\Gamma_{t001}\}\Psi_{t0} \right. \\
&\left. \quad + \{\Lambda_{t10}\Gamma_{t110} + (1-\Lambda_{t10})\Gamma_{t100} - \Lambda_{t11}\Gamma_{t111} - (1-\Lambda_{t11})\Gamma_{t101}\}\Psi_{t1}\right],
\end{aligned}
$$

*where* $\Lambda_{tmv} = \Pr(M_{i2} = 1 \mid T_i = t, M_{i1} = m, V_i = v)$, $\Gamma_{tm_1m_2v} = \mathbb{E}(Y_{i2} \mid T_i = t, M_{i1} = m_1, M_{i2} = m_2, V_i = v)$, *and* $\Psi_{tm} = \Pr(M_{i1} = m \mid T_i = t)/\{(\Lambda_{t00} - \Lambda_{t01})\Pr(M_{i1} = 0 \mid T_i = t) + (\Lambda_{t11} - \Lambda_{t10})\Pr(M_{i1} = 1 \mid T_i = t)\}$.

We begin by defining a trichotomous variable $L_{it} \in \{-1, 0, 1\}$ to indicate the *pliability* type of unit $i$ with respect to the encouragement $V_i$ under treatment status $t$. That is, those units with $L_{it} = 0$ are *pliable* in the sense that their mediator variable always takes the value as encouraged, i.e., $M_{i2}(1-t, 0) = 1 - m$ and $M_{i2}(1-t, 1) = m$ given $M_{i1} = m$ and $T_i = t$. For those with $L_{it} = 1$, the value of the mediator in the second experiment is always the same as in the first experiment, i.e., $M_{i2}(1-t, 0) = M_{i2}(1-t, 1) = m$, whereas for those with $L_{it} = -1$ the second mediator status is the opposite of the first mediator status regardless of the encouragement, i.e., $M_{i2}(1-t, 0) = M_{i2}(1-t, 1) = 1 - m$. By Assumption 12, these three types exhaust all the possible pliability types, and the latter two constitute the group of *non-pliables* in this population.

Next, note that the proportion of each pliability type is identifiable for each stratum defined by $T_i$ and $M_{i1}(t)$ because of the randomization of encouragement. That is, we have the following equalities,

$$
\begin{aligned}
\phi_{1tm} &= \Pr(M_{i2} = m \mid T_i = t, M_{i1} = m, V_i = 0), \\
\phi_{-1tm} &= \Pr(M_{i2} = 1 - m \mid T_i = t, M_{i1} = m, V_i = 1), \\
\phi_{0tm} &= \Pr(M_{i2} = m \mid T_i = t, M_{i1} = m, V_i = 1) - \phi_{1tm} \\
&= \Pr(M_{i2} = 1 - m \mid T_i = t, M_{i1} = m, V_i = 0) - \phi_{-1tm},
\end{aligned}
$$

for $t, m = 0, 1$, where $\phi_{ptm} = \Pr(L_{it} = p \mid T_i = t, M_{i1}(t) = m)$. In addition, we also have the following equalities,

$$
\begin{aligned}
\Gamma_{tm(1-m)1} &= \mathbb{E}(Y_i(1-t, 1-m) \mid T_i = t, M_{i1} = m, L_{it} = -1) \\
\Gamma_{tmm0} &= \mathbb{E}(Y_i(1-t, m) \mid T_i = t, M_{i1} = m, L_{it} = 1) \\
\Gamma_{tm(1-m)0} &= \frac{\phi_{-1tm}}{1 - \phi_{1tm}} \mathbb{E}(Y_i(1-t, 1-m) \mid T_i = t, M_{i1} = m, L_{it} = -1) \\
&\quad + \frac{\phi_{0tm}}{1 - \phi_{1tm}} \mathbb{E}(Y_i(1-t, 1-m) \mid T_i = t, M_{i1} = m, L_{it} = 0) \\
\Gamma_{tmm1} &= \frac{\phi_{0tm}}{1 - \phi_{-1tm}} \mathbb{E}(Y_i(1-t, m) \mid T_i = t, M_{i1} = m, L_{it} = 0) \\
&\quad + \frac{\phi_{1tm}}{1 - \phi_{-1tm}} \mathbb{E}(Y_i(1-t, m) \mid T_i = t, M_{i1} = m, L_{it} = 1)
\end{aligned}
$$

for any $t, m = 0, 1$. By solving this system of equations, we can identify all the conditional expectations in the above expression. Then, the average indirect effects for the pliable group can be identified using the following relationships,

$$
\begin{aligned}
&\mathbb{E}(Y_{1i}(1-t, M_{1i}(1-t)) \mid L_{it} = 0, V_i = 0, T_i = t) \\
&= \mathbb{E}(Y_{i2}(1-t, M_{i2}(1-t, 0), 0) \mid L_{it} = 0, V_i = 0, T_i = t) \\
&= \sum_{m=0,1} \mathbb{E}(Y_{i2}(1-t, M_{i2}(1-t, 0), 0) \mid M_{i1} = m, L_{it} = 0, V_i = 0, T_i = t) \\
&\qquad\qquad \times \Pr(M_{i1} = m \mid L_{it} = 0, V_i = 0, T_i = t) \\
&= \sum_{m=0,1} \mathbb{E}(Y_i(1-t, 1-m) \mid M_{i1} = m, L_{it} = 0, T_i = t) \times \Pr(M_{i1} = m \mid L_{it} = 0, T_i = t),
\end{aligned}
$$

where the first equality follows from Assumption 11, the second from the law of total expectations, and the third from Assumptions 11 and 13 as well as the definition of the pliability types. In addition, we have

$$
\begin{aligned}
&\mathbb{E}(Y_{1i}(1-t, M_{1i}(t)) \mid L_{it} = 0, V_i = 1, T_i = t) \\
&= \mathbb{E}(Y_{i2}(1-t, M_{i2}(1-t, 1), 1) \mid L_{it} = 0, V_i = 1, T_i = t) \\
&= \sum_{m=0,1} \mathbb{E}(Y_{i2}(1-t, M_{i2}(1-t, 1), 1) \mid M_{i1} = m, L_{it} = 0, V_i = 1, T_i = t) \\
&\qquad\qquad \times \Pr(M_{i1} = m \mid L_{it} = 0, V_i = 1, T_i = t) \\
&= \sum_{m=0,1} \mathbb{E}(Y_i(1-t, m) \mid M_{i1} = m, L_{it} = 0, T_i = t) \times \Pr(M_{i1} = m \mid L_{it} = 0, T_i = t),
\end{aligned}
$$

where the first equality follows from Assumption 11 and the definition of the pliability types, the second from the law of total expectation and the third from Assumptions 11, 13 and the definition of $L_{it}$. Note

that the marginal proportion of pliable units is given by,

$$\Pr(L_{it} = 0) \quad = \quad \sum_{m=0,1} \phi_{0tm} \Pr(T_i = t, M_{i1} = m)$$

Then, we can use Bayes' rule to obtain $\Pr(M_{i1} = m \mid L_{it} = 0, T_i = t) = \phi_{0tm} \Pr(T_i = t, M_{i1} = m)/\Pr(L_{it} = 0)$. Finally, expressions given in Theorem 3 can be obtained by substituting observed quantities into the above expressions. $\qquad\square$

# References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, **91**(434), 444–455.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, **92**, 1171–1176.

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51**(6), 1173–1182.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal?: A field experiment on labor market discrimination. *American Economic Review*, **94**(4), 991–1013.

Brader, T., Valentino, N., and Suhay, E. (2008). What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat. *American Journal of Political Science*, **52**(4), 959–978.

Bullock, J., Green, D., and Ha, S. (2010). Yes, But What's the Mechanism? (Don't Expect an Easy Answer). *Journal of Personality and Social Psychology*, **98**(4), 550–558.

Camerer, C., Loewenstein, G., and Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, **43**, 9–64.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, **24**(3), 175–199.

Cox, D. R. (1958). *Planning of Experiments*. John Wiley & Sons, New York.

Dawid, A. (2000). Causal Inference without Counterfactuals. *Journal of the American Statistical Association*, **95**(450).

Deaton, A. (2009). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. *Proceedings of the British Academy*, **162**, 123–160.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, London.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(2), 199–215.

Hainmueller, J. and Hiscox, M. J. (2010). Attitudes toward highly skilled and low-skilled immigration: Evidence from a survey experiment. *American Political Science Review*, **104**(01), 61–84.

Heckman, J. J. and Smith, J. A. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives*, **9**(2), 85–110.

Hedström, P. and Ylikoski, P. (2010). Causal Mechanisms in the Social Sciences. *Annual Review of Sociology*, **36**, 49–67.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, **81**, 945–960.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, **18**, 449–84.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, **15**(4), 309–334.

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, **25**(1), 51–71.

Imai, K., Tingley, D., and Yamamoto, T. (2011a). Replication data for: Experimental designs for identifying causal mechanisms. hdl:1902.1/16416. The Dataverse Network.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011b). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, **105**(4), 765–789.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods*, **13**(4), 314–336.

Jones, B. and Kenward, M. G. (2003). *Design and Analysis of Cross-over Trials*. Chapman & Hall, London, 2nd edition.

Kaufman, S., Kaufman, J. S., and MacLehose, R. F. (2009). Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. *Journal of Statistical Planning and Inference*, **139**, 3473–3487.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., and Fehr, E. (2006). Diminishing Reciprocal Fairness by Disrupting the Right Prefrontal Cortex. *Science*, **314**(5800), 829–832.

Lerner, J. S. and Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*, **81**(1), 146–159.

Little, D. (1990). *Varieties of social explanation: An introduction to the philosophy of social science*. Westview Press, Boulder, Co.

Ludwig, J., Kling, J. R., and Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives*, page Forthcoming.

Mackie, J. (1965). Causes and conditions. *American Philosophical Quarterly*, **2**(4), 245–264.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Harvard University Press.

Martin, A. and Gotts, S. J. (2005). Making the causal link: frontal cortex activity and repetition priming. *Nature Neuroscience*, **8**, 1134–1135.

Mattei, A. and Mealli, F. (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **73**(5), 729–752.

Paus, T. (2005). Inferring causality in brain images: a perturbation approach. *Philosophical Transactions B*, **360**(1457), 1109.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420, San Francisco, CA. Morgan Kaufmann.

Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, **17**(3), 276–284.

Robertson, E. M., Théoret, H., and Pascual-leone, A. (2003). Studies in cognition: The problems solved and created by transcranial magnetic stimulation. *J. Cognitive Neuroscience*, **15**(7), 948–960.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems (eds., P.J. Green, N.L. Hjort, and S. Richardson)*, pages 70–81. Oxford University Press, Oxford.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**(2), 143–155.

Rothman, K. and Greenland, S. (2005). Causation and causal inference in epidemiology. *American Journal of Public Health*, **95**(S1), S144.

Rothman, K. J. (1976). Causes. *American Journal of Epidemiology*, **104**(6), 587–592.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes (with discussions). *Scandinavian Journal of Statistics*, **31**(2), 161–170.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2003). The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science*, **300**(5626), 1755–1758.

Shpitser, I. and VanderWeele, T. J. (2011). A complete graphical criterion for the adjustment formula in mediation analysis. *International Journal of Biostatistics*, **7**(1), Article 16.

Sjölander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine*, **28**(4), 558–571.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, **33**(2), 230–251.

Spencer, S., Zanna, M., and Fong, G. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, **89**(6), 845–851.

VanderWeele, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statistics & Probability Letters*, **78**, 2957–2962.

VanderWeele, T. J. (2009). Mediation and mechanism. *European Journal of Epidemiology*, **24**, 217–224.

VanderWeele, T. J. and Robins, J. M. (2007). The identification of synergism in the sufficient-component-cause framework. *Epidemiology*, **18**(3), 329–339.

VanderWeele, T. J. and Robins, J. M. (2009). Minimal sufficient causation and directed acyclic graphs. *Annals of Statistics*, **37**(3), 1437–1465.