Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2013)
176, *Part* 1, *pp.* 5–51

Experimental designs for identifying causal mechanisms

Kosuke Imai,

Princeton University, USA

Dustin Tingley

Harvard University, Cambridge, USA

and Teppei Yamamoto

Massachusetts Institute of Technology, Cambridge, USA

[*Read before* The Royal Statistical Society *on Wednesday, March 14th, 2012, the President*, Professor V. S. Isham, *in the Chair*]

Summary. Experimentation is a powerful methodology that enables scientists to establish causal claims empirically. 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 strives to identify causal mechanisms. We consider several experimental designs that help to identify average natural indirect effects. Some of these designs require the perfect manipulation of an intermediate variable, whereas 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 designs proposed.

Keywords: Causal inference; Direct and indirect effects; Identification; Instrumental variables; Mediation

1. Introduction

Over the last century and across numerous disciplines, experimentation has been a powerful methodology to test scientific theories. As Neyman demonstrated in 1923 (see Neyman (1990)), 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, whereas experiments can identify average causal effects, they cannot explain causal mechanisms (e.g. Heckman and Smith (1995), Cook (2002) and 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,

Address for correspondence: Kosuke Imai, Department of Politics, Princeton University, Princeton, NJ 08544, USA.
E-mail: kimai@princeton.edu

 0964–1998/13/176005

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 a '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) and Imai, Keele and Yamamoto (2010); 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 that are 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 whereas 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 on the basis of the treatment and mediator values in the previous experiment. We show that the two designs have a potential to improve the identification power of the single-experiment design significantly.

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 manipulate the mediator perfectly. 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 the 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 a strong 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 has been 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 we 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) and VanderWeele and Robins (2009)), as well as other related quantities that have appeared in recent works (Geneletti, 2007; Spencer *et al.*, 2005).

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) and Little (1990)). Specifically, we study the identification of a simple causal mechanism, which is represented by the full arrows in the causal diagram of Fig. 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 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 define indirect effects formally 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 are 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 observe only one of the potential outcomes, i.e. $Y_i = Y_i\{T_i, M_i(T_i)\}$. Throughout this paper, we assume no interference between 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 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.

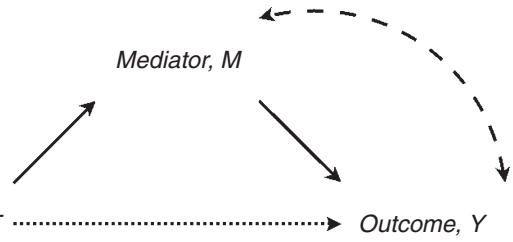

Fig. 1. Diagram for a simple causal mechanism: ⟶, causal mechanism of interest where the causal effect of the treatment on the outcome is transmitted through the intermediate variable or the mediator; ⋯⋯➤, all the other possible causal mechanisms; - - - -➤, possible presence of confounders between the mediator and outcome, which typically cannot be ruled out in the single-experiment design

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

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

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

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

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 we change 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 those 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

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

for $t = 0, 1$. The key counterfactual question is: what difference in the outcome would result if we change 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 Rubin (1974) and Holland (1986), the fundamental problem of causal inference under the potential outcomes framework is that given any unit we 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, whereas $Y_i\{t, M_i(t)\}$ 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 $\bar{\tau} \equiv \mathbb{E}[Y_i\{1, M_i(1)\} - Y_i\{0, M_i(0)\}]$, $\bar{\delta}(t) \equiv \mathbb{E}[Y_i\{t, M_i(1)\} - Y_i\{t, M_i(0)\}]$ and $\bar{\zeta}(t) \equiv \mathbb{E}[Y_i\{1, M_i(t)\} - Y_i\{0, M_i(t)\}]$, for $t = 0, 1$.

Altogether, 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. The fact that we can decompose the average causal effect $\bar{\tau}$ into the sum of average direct and indirect effects 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 Fig. 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 recently been formalized under the sufficient cause framework (e.g. Rothman and Greenland (2005) and VanderWeele and Robins (2007)). VanderWeele (2009) formally studied the relationship of this alternative definition to the process-based definition of the causal mechanism by 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 is 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. VanderWeele and Robins (2008) derived the conditions under which this procedure is justified.

Although justifiable for analysing causal components, such a procedure is generally not useful for the study of causal processes. For example, whereas 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

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

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 we must assume the *absence* of interactions to identify causal processes.

Finally, some advocate the alternative definitions of direct and indirect effects based on principal stratification (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 indirect 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 VanderWeele (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 set-up that was reviewed above, we study the single-experiment design, which is the most common experimental design employed by applied researchers 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.

2.3.1. Set-up

The randomization of the treatment (possibly conditional on a vector of observed pretreatment 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. Fig. 1 encodes this assumption since no broken bidirectional arrow is depicted between T and M or T and Y. Formally, this can be written as follows.

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.

2.3.2. *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, Keele and Yamamoto (2010) showed 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 follows.

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 non-parametrically identified (see theorem 1 of Imai, Keele and Yamamoto (2010)). Under a linearity assumption, this assumption also justifies the common method that was popularized by Baron and Kenny (1986) (see Imai, Keele and Tingley (2010)). The discussion of other assumptions that are 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 that 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 are unobserved confounders that affect both the mediator and the outcome. Imai, Keele and Yamamoto (2010) also pointed out that, whereas observed *pretreatment* confounders of the relationship between the mediator and outcome can be controlled for in straightforward ways, the mediator–outcome confounders that are *post treatment* 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.

2.3.3. *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) and 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 illustration, we assume that both the outcome and the mediator are binary. Then, it is straightforward to obtain the sharp bounds by 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 0 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 evaluate formally how robust one's conclusions are in the presence of possible violations of a key identifying assumption (see Imai, Keele and Tingley (2010) and Imai, Keele and Yamamoto (2010)).

2.3.4. Example

Brader *et al.* (2008) examined 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 2×2 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 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. Brader *et al.* (2008) 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, compromising the credibility of causal mediation analysis. In many psychological experiments including this one, researchers are interested in psychological mechanisms that explain behavioural 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 the strong assumption of sequential ignorability.

3. Experimental designs with direct manipulation

Many critics of the single-experiment design view the randomization of the 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) and 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. Although 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 homogeneous across units, we 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, Keele, Tingley and Yamamoto (2011)).

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. 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 that was 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 considered an identification strategy under such a design. Our identification analysis differs from Pearl's in that he considered identification under a sequential ignorability assumption that was similar to assumption 2. We also derive the sharp bounds on the average indirect effects in the absence of any assumption that is not justified by the design itself.

3.1.1. Set-up

Suppose that we use $D_i = 0$ and $D_i = 1$ to indicate that unit i belongs to the first and second experiment respectively. 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 $m \in \mathcal{M}$,

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

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 $m \in \mathcal{M}$,

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

3.1.2. 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. In contrast, $\mathbb{E}[Y_i\{t, M_i(t')\}]$ is not identified as the following decomposition shows:

$$\mathbb{E}[Y_i\{t, M_i(t')\}] = \int \mathbb{E}\{Y_i(t, m) \mid M_i(t') = m\} \, \mathrm{d}F_{M_i|T_i=t', D_i=0}(m), \tag{5}$$

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) can also be unbounded.

To achieve identification, we may rely on the assumption that there is 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, a point to which we return shortly below.

The following theorem shows that, if we are 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, 4 and 5, for $t = 0, 1$, the average indirect effects are identified and given by

$$\begin{aligned}\bar{\delta}(t) = {} & \mathbb{E}(Y_i \mid T_i = 1, D_i = 0) - \mathbb{E}(Y_i \mid T_i = 0, D_i = 0) \\ & - \int \{\mathbb{E}(Y_i \mid T_i = 1, M_i = m, D_i = 1) - \mathbb{E}(Y_i \mid T_i = 0, M_i = m, D_i = 1)\} \, \mathrm{d}F_{M_i|D_i=1}(m).\end{aligned}$$

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. Theorem 1 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 two terms of the expression in theorem 1 are 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 remaining 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 by 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 other 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 on the basis of theorem 1, however, will lead to a severely biased estimate: the estimate converges to $\bar{\delta}(t) = 0$ whereas 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, i.e. the average indirect effect depends on 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 observe only 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 pretreatment characteristics that are known to be related to the magnitude of interaction effects and to implement the parallel design within each stratum defined by these pretreatment variables. Alternatively, a sensitivity analysis such as that developed by Imai and Yamamoto (2012) can be used to examine the robustness of empirical findings to the violation of this assumption.

3.1.3. 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 power of identification 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 to that of 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 is a trade-off 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.

3.1.4. Example

In behavioural neuroscience, scholars have used brain imaging technology, such as functional magnetic resonance imaging, to measure the operation of neural mechanisms. Functional magnetic resonance imaging 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 to manipulate activation of the region. This allows in principle for a direct manipulation of the hypothesized neural mechanism linking a stimulus with a behavioural response. A growing number of studies use TMS (e.g. Martin and Gotts (2005) and Paus (2005)) because it 'directly leads to causal inferences about brain functioning rather than the purely associational evidence provided by imaging techniques' (Camerer *et al.* (2005), pages 13–14).

For example, Knoch *et al.* (2006) used TMS to understand the neural mechanisms underlying common behaviour that is 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 evidence based on functional magnetic resonance imaging 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. Given this result, Knoch *et al.* (2006) used TMS to investigate whether the activity in the dorsolateral prefrontal cortex controls an impulse to reject unfair offers or regulates a selfish impulse. It was argued that, if the dorsolateral prefrontal cortex were to be deactivated and individuals accept more unfair offers, then this would represent evidence that the dorsolateral prefrontal cortex serves the role of implementing fair behaviour 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 dorsolateral prefrontal cortex. 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 experiment of Knoch *et al.* (2006)? Their results suggest not. They found 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 dorsolateral prefrontal cortex causally interact in determining subjects' behaviour, this does not necessarily imply that the dorsolateral prefrontal cortex represents a causal process through which the effect of the fairness treatment is transmitted.

3.2. Crossover design

To improve further on the parallel design, we must directly address the fundamental problem

of identifying causal mechanisms that was discussed in Section 2. For example, we can never observe $M_i(1)$ for units with $T_i = 0$, but we must identify $\mathbb{E}[Y_i\{0, M_i(1)\}]$ to identify $\bar{\delta}(0)$. Here, we consider the *crossover design* where each experimental unit is exposed to both 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 carry-over 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 that suggested by Robins and Greenland (1992) where 'both exposure and the cofactor intervention [i.e. mediator manipulation] are randomly assigned in both time periods' (page 153). They showed 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 that is identical to our crossover design was also mentioned by Pearl (2001), page 1574, albeit only in passing.

3.2.1. Set-up

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, i.e. 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 a carry-over effect as is often done in the standard crossover trials. Specifically, we assume that the treatment that is administered in the first period does not affect the average outcome in the second period, as well as that there is no period effect (i.e. 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 carry-over effects under the crossover design).

$$\mathbb{E}[Y_{i1}\{t, M_i(t)\}] = \mathbb{E}\{Y_{i2}(t, m)\} \qquad \text{if } M_i(t) = m,$$

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. 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 test assumption 7 directly; however, the assumption can be partially tested if we modify the second experiment to include an optional subgroup for each treatment group which does not receive any mediator manipulation. This test can be done by comparing the average observed outcome among each of these subgroups with the average outcome among the opposite treatment group in the first experiment. If the difference between these values is insignificant for both treatment conditions, the analyst can know that the no-carry-over effect (but not necessarily the consistency) assumption is plausible.

3.2.2. Identification

Under the crossover design, experimenters attempt to measure potential outcomes under different treatment and mediator values for each unit. This helps to address the fundamental problem of identifying causal mechanisms that was discussed in Section 2. The following theorem summarizes the fact that under the crossover design the randomization of the treatment and the assumption of consistency and no carry-over 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

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

A proof is straightforward, and therefore it is omitted.

3.2.3. Sharp bounds

Under the crossover design, the assumption of consistency and no carry-over 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 carry-over effect plays an essential role although the difference is that under the standard crossover design this assumption can be directly tested.

3.2.4. Example

In a landmark paper, Bertrand and Mullainathan (2004) conducted a randomized field experiment to test labour market discrimination against African Americans. They created fictitious

résumés, 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 potentially keeping their perceived qualifications (the mediator M_i) constant. These *résumés* are then randomly sent to potential employers and callback rates for interviews are measured as the outcome variable of interest. Bertrand and Mullainathan (2004) found that the *résumés* with white names are more likely to yield callbacks than those with black names.

Under the original experimental design, the researchers could 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 *résumés*. 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 perceived qualifications.

The identification of this quantity is useful to isolate the degree to which African American job applicants are discriminated not on the basis of qualifications but on their race. If the quantity is positive, then it may suggest racial discrimination in the labour market. The key difference between the two quantities is that the former is conditional on a particular qualification m assigned by experimentalists whereas 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 *résumés* (i.e. setting M_i to some arbitrary m). Instead, we could sample actual *résumés* of African American job applicants to obtain $M_i(0)$. Sending these *résumés* 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 to identify the counterfactual outcome $\mathbb{E}[Y_i\{1, M_i(0)\} \mid T_i = 0]$ without changing the other parts of the *résumés* (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-carry-over effect assumption may be problematic if the same *résumé* with different names is sent to the same potential employer over two time periods. Fortunately, this problem can be overcome by sending these *résumés* 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 carry-over effects only need to hold in expectation. Thus, researchers will be able to infer how much of labour 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 a strong 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. 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) and 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. 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.

4.1.1. Set-up

Formally, let Z_i represent the ternary encouragement variable where it is equal to 1 or -1 if subject i is respectively positively or 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 the 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 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 $m = 0, 1$,

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

Here, both the treatment and the encouragement are assumed to be under perfect control of the analyst and thus conditioning on pretreatment 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 that 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, i.e. there are no 'defiers' who behave exactly oppositely to the encouragement. Without loss of generality, these two assumptions can be formalized as follows.

Assumption 9 (consistency under the parallel encouragement design). For all $t = 0, 1$ and $z, z' = -1, 0, 1$,

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

Assumption 10 (monotonicity). For $t = 0, 1$,

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

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. The quantity that 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 $t \neq t'$.

4.1.2. Sharp bounds

Given this set-up, we study the identification power of the parallel encouragement design again using a bounds approach. Again, for simplicity and comparison with the other designs, we focus on the situation where the outcome is also binary. In this case, under assumptions 8–10, the sharp bounds can be derived numerically by 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.

4.1.3. 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. Although this prevents researchers from using the parallel design, the parallel encouragement design may be applicable to this type of psychological experiment. 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 are several ways to manipulate emotion indirectly. 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 to make the consistency assumption more plausible.

The other key assumption of monotonicity (assumption 10) will be violated if there are any subjects whose level of anxiety 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. Crossover encouragement design

It is also possible to generalize the crossover design that was described in Section 3.2 so that the imperfect manipulation of the mediator is allowed. Under this *crossover encouragement design*, after the treatment has been 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 that we call the *pliable* units. Whereas 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 behaviour in response to the encouragement. We then combine this with other information obtained from the second period to identify causal mechanisms among the pliables.

4.2.1. Set-up

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 in 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 carry-over effect (for both the mediator and the outcome), can be formalized as follows.

Assumption 11 (consistency and no carry-over effects under crossover encouragement design). For all $t, t', v = 0, 1$,

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

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 carry-over effects

assumptions. First, it is possible to test whether the first equality in assumption 11 holds on average by comparing $\mathbb{E}(M_{i1}|T_i = t)$ with $\mathbb{E}(M_{i2}|T_i = 1-t, V_i = 0)$ 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 the encouragement are randomized. Second, the second equality in assumption 11 can be partially tested by comparing $\mathbb{E}(Y_{i1}|T_i = t, M_{i1} = m)$ with $\mathbb{E}(Y_{i2}|T_i = 1-t, M_{i2} = m, V_i = 0)$ for $m = 0, 1$. This is because these two quantities are equal to $\mathbb{E}\{Y_{i1}(t,m)|M_{i1}(t) = m\}$ and $\mathbb{E}\{Y_{i2}(t,m,0)|M_{i2}(t,0) = m\}$ 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 measuring the first period outcome because it allows testing whether there is any carry-over effect and it often involves little additional cost.

In addition to these assumptions, which are essentially equivalent to the assumptions that are made under the crossover design, we rely on 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 follows.

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.

Assumption 13 (randomization of treatment and encouragement). For any $m^* \in \mathcal{M}$ and $t^* \in \{0, 1\}$,

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

4.2.2. *Identification*

Under these assumptions and binary mediator and outcome variables, we 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.

4.2.3. *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. They studied how immigration preferences of low income US citizens are

influenced by whether they are asked about high or low skill immigrants. One of the hypotheses that they considered 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 effect of immigrants on access to public services. Although Hainmueller and Hiscox (2010) could not directly test this mechanism, a modification of their original experimental design may permit this.

The study used the standard 2×2 crossover design where survey respondents were first randomly asked to consider either high or low skill immigrants and then to express their policy preferences about increasing immigration. 2 weeks later, the same respondents were surveyed again, except that they were asked about the other skill group, thereby reversing the treatment. Hainmueller and Hiscox (2010) 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 favourable 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-carry-over 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 2 weeks and found little carry-over effects. The long wash-out period in their design makes the no-carry-over effect 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. Numerical example

We now illustrate some of our analytical results by 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 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 0.

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

Table 1. Descriptive statistics and estimated average treatment effects from the immigration experiment†

Response variable	*Results for treatment group*		*Results for control group*		*Average treatment effect (standard error)*
	Mean	*Standard deviation*	*Mean*	*Standard deviation*	
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		

†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.

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 on the basis of these two variables.

Fig. 2 presents the sharp bounds on the average indirect effects for $t = 1$ (Fig. 2(a)) and $t = 0$ (Fig. 2(b)) under different experimental designs calculated from the simulated population. In both panels, the top three full 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 both $\bar{\delta}(1)$ and $\bar{\delta}(0)$ and include 0 ($[-0.175, 0.825]$ 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 0.

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 0 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 is the standard single-experiment design.

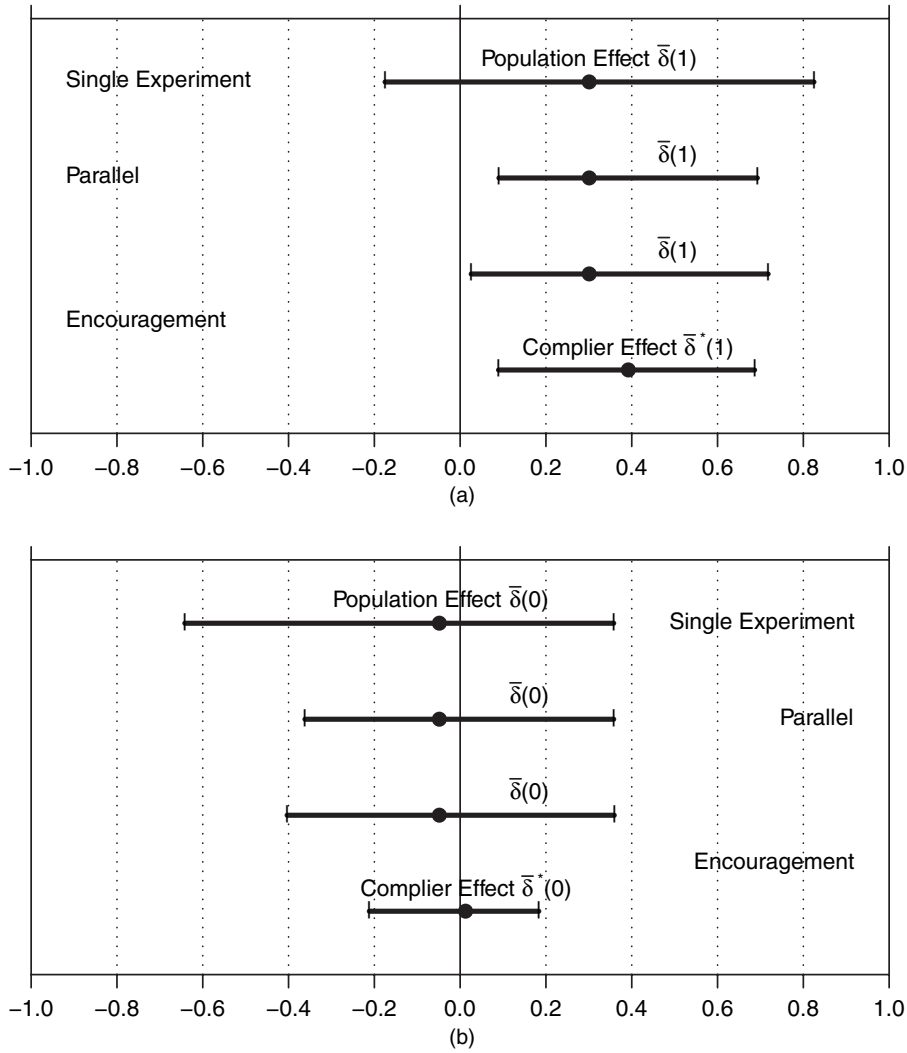

Fig. 2. Identification power of alternative experimental designs: sharp bounds on the average indirect effects for (a) the treatment and (b) the control conditions calculated on the basis of the hypothetical population that we generated (●, true values of $\bar{\delta}(t)$ (top three) and $\bar{\delta}^{*}(t)$ (bottom); ⟼⟻ 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 with the traditional 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 often been 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 to test the existence of hypothesized causal mechanisms empirically. In this paper, we answer this question by proposing alternative experimental designs and analysing 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 that we consider require the direct manipulation of mediator whereas 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 that was discussed in Section 3.1) as well as the creativity on the part of experimenters (e.g. the labour market discrimination experiment that was discussed in Section 3.2).

The methodology proposed emphasizes the identification assumptions that are directly linked to experimental design rather than those on the characteristics of experimental units. Although experimenters can play only 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 identify causal mechanisms experimentally through clever manipulations and future technological developments. Although 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.

Acknowledgements

Replication materials for this paper are available on line as Imai, Tingley and Yamamoto (2011). We thank Erin Hartman and Adam Glynn as well as seminar participants at Columbia University (Political Science), the University of California at Berkeley (Biostatistics), the University of Maryland Baltimore County (Mathematics and Statistics) and New York University (the Mid-Atlantic Causal Inference Conference) for helpful comments. Detailed suggestions from four referees and the Editor significantly improved the presentation of this paper. Financial support from National Science Foundation grants SES–0849715 and SES–0918968 is acknowledged.

Appendix A

A.1. Relation to Geneletti's (2007) indirect effects

On the basis of a non-counterfactual framework of causal inference, Geneletti (2007) showed 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 re-expressed by 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 comparison can be rewritten as $\mathbb{E}\{Y_i(t, M_1) - Y_i(t, M_0) | M_0 = M_i(0), M_1 = M_i(1)\}$.

The difference between $\bar{\delta}^{\dagger}(t)$ and $\bar{\delta}(t)$ is rather subtle but important. For illustration, we use Geneletti's example (see her section 3.1.2, point (b)) about a drug treatment for a particular disease that may trigger headaches as a 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 (which is a difficult assumption to maintain given the heterogeneity of human populations). Our approach, however, avoids this exchangeability assumption and develops experimental designs that help to 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 that are implied by assumption 1 alone. Here we take a simpler alternative approach which uses the equality given in Section 2.1, $\bar{\tau} = \bar{\delta}(t) + \bar{\zeta}(1-t)$, i.e. 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 $t = 0, 1$:

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

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

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 $\int \mathbb{E}(Y_i \mid T_i = t, X_i = x, D_i = 0)\, \mathrm{d}F_{X_i \mid D_i = 0}(x)$ and the latter from the second experiment by $\int \mathbb{E}(Y_i \mid T_i = t, M_i = m, X_i = x, D_i = 1)\, \mathrm{d}F_{X_i \mid D_i = 1}(x)$. 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 $m \in \mathcal{M}$.

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

$$\mathbb{E}[Y_i\{1, M_i(0)\}] = \sum_{y=0}^{1} \sum_{m=0}^{1} (\pi_{1ym1} + \pi_{y1m0})$$

where $\pi_{y_1 y_0 m_1 m_0} = \Pr\{Y_i(1, 1) = y_1, Y_i(1, 0) = y_0, M_i(1) = m_1, M_i(0) = m_0\} \geqslant 0$ 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:

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

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

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

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

The sharp bounds on $\bar{\delta}(1)$ can then be obtained by combining these constraints 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

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

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

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

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 $P_{001} \leqslant Q_{001} \leqslant P_{001} + P_{011} + P_{111}$, $P_{011} \leqslant Q_{011} \leqslant P_{001} + P_{101} + P_{011}$ and $-P_{101} - P_{111} \leqslant Q_{001} - Q_{111} \leqslant P_{001} + P_{011}$. 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 expressions 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 is a trade-off 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 or small, then the lower bound of $\bar{\delta}(1)$ will be respectively large or 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

$$\delta^*(t) = \frac{\sum_{m'_{-1}=0}^{1} \sum_{m'_1=0}^{1} (\psi^{01}_{m'_{-1} 0 m'_1 011} + \psi^{10}_{m'_{-1} 1 m'_1 001} - \psi^{01}_{m'_{-1} 1 m'_1 001} - \psi^{10}_{m'_{-1} 0 m'_1 011})}{\Pr\{M_i(t, -1) = 0, M_i(t, 0) = M_i(t, 1) = 1\} + \Pr\{M_i(t, -1) = M_i(t, 0) = 0, M_i(t, 1) = 1\}}, \quad (10)$$

where

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

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 $P^{\dagger}_{ymt}=\Pr(Y_i=y, M_i=m|T_i=t, Z_i=-1)$ and $P^{*}_{ymt}=\Pr(Y_i=y, M_i=m|T_i=t, Z_i=1)$. 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 by using the above framework.

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 and 13, the average indirect effect among the pliable units is identified and given by

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

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

$$\Psi_{tm}=\frac{\Pr(M_{i1}=m|T_i=t)}{(\Lambda_{t00}-\Lambda_{t01})\Pr(M_{i1}=0|T_i=t)+(\Lambda_{t11}-\Lambda_{t10})\Pr(M_{i1}=1|T_i=t)}.$$

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-pliable* units 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, i.e. we have the equalities

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

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 equalities

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

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 by using the following relationships:

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

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

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

where the first equality follows from assumption 11 and the definition of the pliability types, the second from the law of toal expectation and the third from assumptions 11 and 13 and the definition of L_{it}. Note that the marginal proportion of pliable units is given by

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

Then, we can use the Bayes rule to obtain $\Pr(M_{i1} = m | L_{it} = 0, T_i = t) = \phi_{0tm} \Pr(T_i = t, M_{i1} = m_1) / \Pr(L_{it} = 0)$. Finally, expressions given in theorem 3 can be obtained by substituting observed quantities into the above expressions.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Ass.*, **91**, 444–455.

Balke, A. and Pearl, J. (1997) Bounds on treatment effects from studies with imperfect compliance. *J. Am. Statist. Ass.*, **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. *J. Personlty Socl Psychol.*, **51**, 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. *Am. Econ. Rev.*, **94**, 991–1013.

Brader, T., Valentino, N. and Suhay, E. (2008) What triggers public opposition to immigration? anxiety, group cues, and immigration threat. *Am. J. Polit. Sci.*, **52**, 959–978.

Bullock, J., Green, D. and Ha, S. (2010) Yes, but what's the mechanism? (don't expect an easy answer). *J. Personlty Socl Psychol.*, **98**, 550–558.
Camerer, C., Loewenstein, G. and Prelec, D. (2005) Neuroeconomics: how neuroscience can inform economics. *J. Econ. Lit.*, **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. *Educ. Evaln Poly Anal.*, **24**, 175–199.
Cox, D. R. (1958) *Planning of Experiments*. New York: Wiley.
Dawid, A. (2000) Causal inference without counterfactuals. *J. Am. Statist. Ass.*, **95**, 407–448.
Deaton, A. (2009) Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development. *Proc. Br. Acad.*, **162**, 123–160.
Geneletti, S. (2007) Identifying direct and indirect effects in a non-counterfactual framework. *J. R. Statist. Soc.* B, **69**, 199–215.
Hainmueller, J. and Hiscox, M. J. (2010) Attitudes toward highly skilled and low-skilled immigration: evidence from a survey experiment. *Am. Polit. Sci. Rev.*, **104**, 61–84.
Heckman, J. J. and Smith, J. A. (1995) Assessing the case for social experiments. *J. Econ. Perspect.*, **9**, 85–110.
Hedström, P. and Ylikoski, P. (2010) Causal mechanisms in the social sciences. *A. Rev. Sociol.*, **36**, 49–67.
Holland, P. W. (1986) Statistics and causal inference (with discussion). *J. Am. Statist. Ass.*, **81**, 945–960.
Holland, P. W. (1988) Causal inference, path analysis, and recursive structural equations models. *Sociol. Methodol.*, **18**, 449–484.
Imai, K., Keele, L. and Tingley, D. (2010) A general approach to causal mediation analysis. *Psychol. Meth.*, **15**, 309–334.
Imai, K., Keele, L., Tingley, D. and Yamamoto, T. (2011) Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. *Am. Polit. Sci. Rev.*, **105**, to be published.
Imai, K., Keele, L. and Yamamoto, T. (2010) Identification, inference, and sensitivity analysis for causal mediation effects. *Statist. Sci.*, **25**, 51–71.
Imai, K., Tingley, D. and Yamamoto, T. (2011) Replication data for: Experimental designs for identifying causal mechanisms. *hdl:1902.1/16416.* Dataverse Network.
Imai, K. and Yamamoto, T. (2012) Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. Submitted to *Polit. Anal.* (Available from `http://imai.princeton.edu/research/medsens.html`.)
Jo, B. (2008) Causal inference in randomized experiments with mediational processes. *Psychol. Meth.*, **13**, 314–336.
Jones, B. and Kenward, M. G. (2003) *Design and Analysis of Cross-over Trials*, 2nd edn. London: Chapman and Hall.
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. *J. Statist. Planng Inf.*, **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**, 829–832.
Lerner, J. S. and Keltner, D. (2001) Fear, anger, and risk. *J. Personlty Socl Psychol.*, **81**, 146–159.
Little, D. (1990) *Varieties of Social Explanation: an Introduction to the Philosophy of Social Science*. Boulder: Westview.
Ludwig, J., Kling, J. R. and Mullainathan, S. (2011) Mechanism experiments and policy evaluations. *J. Econ. Perspect.*, to be published.
Mackie, J. (1965) Causes and conditions. *Am. Philos. Q.*, **2**, 245–264.
Manski, C. F. (1995) *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.
Martin, A. and Gotts, S. J. (2005) Making the causal link: frontal cortex activity and repetition priming. *Nat. Neursci.*, **8**, 1134–1135.
Mattei, A. and Mealli, F. (2011) Augmented designs to assess principal strata direct effects. *J. R. Statist. Soc.* B, **73**, 729–752.
Neyman, J. (1990) On the application of probability theory to agricultural experiments: essay on principles, Section 9. *Statist. Sci.*, **5**, 465–480 (Engl. transl.).
Paus, T. (2005) Inferring causality in brain images: a perturbation approach. *Philos. Trans.* B, **360**, 1109–1114.
Pearl, J. (2001) Direct and indirect effects. In *Proc. 17th Conf. Uncertainty in Artificial Intelligence*, pp. 411–420. San Francisco: Morgan Kaufmann.
Petersen, M. L., Sinisi, S. E. and van der Laan, M. J. (2006) Estimation of direct causal effects. *Epidemiology*, **17**, 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. Cogn. Neursci.*, **15**, 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), pp. 70–81. Oxford: Oxford University Press.
Robins, J. M. and Greenland, S. (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.
Rothman, K. J. (1976) Causes. *Am. J. Epidem.*, **104**, 587–592.

Rothman, K. and Greenland, S. (2005) Causation and causal inference in epidemiology. *Am. J. Publ. Hlth*, **95**, S1, S144.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Rubin, D. B. (2004) Direct and indirect causal effects via potential outcomes (with discussions). *Scand. J. Statist.*, **31**, 161–170.
Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
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**, 1755–1758.
Shpitser, I. and VanderWeele, T. J. (2011) A complete graphical criterion for the adjustment formula in mediation analysis. *Int. J. Biostatist.*, **7**, article 16.
Sjölander, A. (2009) Bounds on natural direct effects in the presence of confounded intermediate variables. *Statist. Med.*, **28**, 558–571.
Sobel, M. E. (2008) Identification of causal parameters in randomized studies with mediating variables. *J. Educ. Behav. Statist.*, **33**, 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. *J. Personlty Socl Psychol.*, **89**, 845–851.
VanderWeele, T. J. (2008) Simple relations between principal stratification and direct and indirect effects. *Statist. Probab. Lett.*, **78**, 2957–2962.
VanderWeele, T. J. (2009) Mediation and mechanism. *Eur. J. Epidem.*, **24**, 217–224.
VanderWeele, T. J. and Robins, J. M. (2007) The identification of synergism in the sufficient-component-cause framework. *Epidemiology*, **18**, 329–339.
VanderWeele, T. J. and Robins, J. M. (2008) Empirical and counterfactual conditions for sufficient cause interactions. *Biometrika*, **91**, 49–61.
VanderWeele, T. J. and Robins, J. M. (2009) Minimal sufficient causation and directed acyclic graphs. *Ann. Statist.*, **37**, 1437–1465.

Discussion on the paper by Imai, Tingley and Yamamoto

Fabrizia Mealli (*Università di Firenze*)

Imai, Tingley and Yamamoto must be congratulated for having attacked the challenging problem of understanding causal mechanisms. I like their engagement in exploring experimental designs, because this task requires spelling out the assumptions that are needed for identification and makes one reflect deeply on the questions that are asked.

Understanding mechanisms is particularly valuable if it helps to design improved interventions. An issue that deserves further discussion is whether investigations on pathways should focus on natural direct and indirect effects. Although these effects have received attention is some disciplines, are they the *natural* estimands that may suggest important pathways? Pearl (2001), paragraph 2.2, originally called them *descriptive* tools for attributing part of the effect of an intervention to an intermediate variable. But I think that the tool often fails to provide a good description of how things work, because of the asymmetric roles of $M_i(0)$ and $M_i(1)$; in general only one of the potential values of the intermediate variable is chosen as descriptive of the *causal forces under natural conditions*. To me, both values are *natural*, in that they describe how an individual reacts to a treatment. Their joint value is essentially a characteristic of a subject, so conceiving a manipulation of one of the two values is like considering changing the value of a pretreatment characteristic. This is essentially why consistency assumptions are rarely credible: they assume that the action that is taken to modify the value of a characteristic of a subject has no consequence on the outcome value. Also, quantities of the type $Y_i\{t, M_i(t')\}$, $t \neq t'$, are ill defined and sometimes difficult to conceive if one is not explicit about the process that led to observing $M_i(0)$ and $M_i(1)$ (Mealli and Rubin, 2003); $Y_i\{t, M_i(t')\}$, $t \neq t'$, are quantities that in a single experiment are '*a priori* counterfactuals' because they cannot be observed for any subset of units. Even assuming that consistency holds, there may be subjects, possibly characterized by covariates' values, for whom a level of M equal to $M_i(0)$ under treatment can never be reached. If this is so, it means that the experiment is seeking an outcome that never occurs in real life, so I fail to understand how such a quantity can have some descriptive power. This suggests that, when interest lies in opening the black box, valuable design issues should be directed more on collecting detailed background covariates and additional outcomes, rather than on generating outcomes under manipulations of the mediating variable.

Despite recognizing the value of the experiments that are proposed by the authors for opening the possibility of identifying causal mechanisms through clever manipulation, I find that the type of settings

where those could be applied most convincingly are those like the example of gender discrimination, where manipulation does not involve human beings directly.

A better description of how things work is provided by looking at the joint value of the natural levels of $M_i(0)$ and $M_i(1)$; those joint values define a stratification of the population into principal strata. Principal strata effects (PSEs), contrasts of $Y(0)$ and $Y(1)$ within principal strata, are well-defined causal quantities, which do not involve *a priori* counterfactuals (Frangakis and Rubin, 2002). If one is seeking information on the effect of an intervention that is not attributable to the change in the intermediate variable, it is sensible to start looking at the effect of the intervention on subjects for whom M *naturally* does not change, i.e. $M_i(0) = M_i(1)$. These PSEs are called dissociative and can be contrasted with associative effects, i.e. effects in principal strata where $M_i(0) \neq M_i(1)$. They allow distinguishing causal effects of T on Y that are associated with causal effects of T on M, from causal effects of T on Y that are dissociative and thereby associated with other causal pathways. The information that is provided by PSEs is extremely valuable: for example, large associative effects relative to small dissociative effects would indicate that the intervention has stronger effects on units where it also has an effect on the mediator. Associative and dissociative effects of equal magnitude would instead indicate that the intervention's effect is the same regardless of whether it has an effect on the mediator, which would suggest some alternative causal pathways through which the intervention has an effect without having an effect on the mediator. Even if these different values of PSEs can be due to principal strata heterogeneity only, an accurate principal strata analysis can provide useful insights on mechanisms and generate useful hypotheses that can be confronted with subject matter knowledge and also tested with a confirmatory experiment on a newly designed intervention. Looking at the distribution of the covariates and outcomes within the strata (Frumento *et al*., 2012) may provide insights on the plausibility of ignorability assumptions for $M(0)$ and $M(1)$ (Jin and Rubin, 2008) to identify the effects of M on Y.

I appreciated the authors' effort to relax perfect manipulation by introducing encouragement designs, providing new alternative approaches to discover causal mechanisms. However, the effects that these designs usually help to reveal are essentially PSEs, and I am glad that the authors recognize their usefulness, despite their *local* nature.

The variety of designs presented by Imai, Tingley and Yamamoto shows how some common jargon is difficult to translate into proper causal statements. They have done a great job by engaging in this challenging area, and it is therefore my pleasure to propose the vote of thanks.

Carlo Berzuini (*University of Cambridge*)

The authors must be congratulated for their very stimulating paper that bridges advanced causal inference methodology and experimental scientific investigation.

The authors choose a causal inference framework based on potential outcomes, where each individual is characterized by a notional value of the response for each possible treatment. These notional values—called potential outcomes—are assumed to be fixed for the individual even before any treatment is applied. In their approach to the encouragement designs, the authors use a powerful device, that of restricting inferential attention to a particular principal stratum (PS), which means a group of individuals defined by the values of two or more potential outcomes for the same variable. A possible difficulty arises here. This is because potential outcomes cannot be jointly observed, and therefore we do not generally know who the individuals in a given PS are. For example, in their treatment of the parallel encouragement design, the authors restrict attention to the PS of individuals characterized by specific patterns of reaction to specific stimuli, a property that we shall not normally be able to check in any given individual. The paper offers examples of clever use of PSs. But the use of this device raises *caveats* that we shall now discuss.

It will suffice to illustrate the issue in relation to the parallel encouragement design, where the authors restrict inferential attention to the PS of compliers, i.e. of those individuals who react to the encouragement in the intended direction ($M(t, 1) = 1, M(t, -1) = 0$). The method assumes that complier status (albeit unobservable) is a fixed and time invariant attribute of an individual. Is this a reasonable assumption? For example, is it reasonable to assume that someone we observe reacting to a specific stimulus with an increase in anxiety will always react to it in the same way? The actual state of affairs might be different. No matter how we circumscribe the problem—the complier status might really remain a random variable, causing individuals to move in and out of the group of compliers in an unpredictable way. In this case our inferences would be based on just those individuals who happened by chance to be compliers during the experiment. Can we, in such a case, claim that we are learning about a stable mechanism of nature? In certain applications, a (real or presumed) natural law (of the kind that we encounter in physics) will

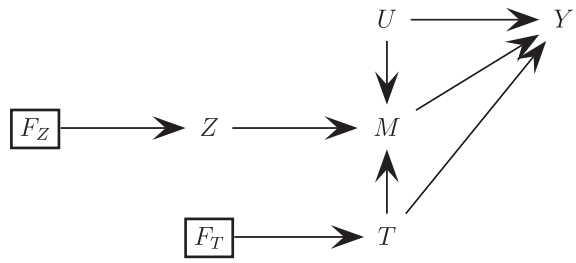

Fig. 3. Causal diagram for the encouragement design, supplemented with intervention indicators (□)

support the claim that the compliers constitute a stable and scientifically meaningful stratum of the population, as illustrated by the application example at the end of these comments.

My next comment concerns a reinterpretation of the counterfactual independences that are used in the paper to express the method assumptions. These independences can be visualized through a causal diagram of the kind shown in Fig. 3. This diagram represents the parallel encouragement design and has been obtained by supplementing the graph of Fig. 1 with a node representing the encouragement variable Z and with the *intervention indicators* F_T and F_Z. The latter are decision variables indicating the manipulation that is performed on the corresponding variable, respectively T or Z, or the absence of such a manipulation, depending on the specific experimental design. In a potential outcomes interpretation of the diagram, U contains the entire collection of potential outcomes for M and Y, and these two variables depend on their direct influences in a deterministic way. With this interpretation, the graph faithfully represents assumptions 8 and 9, which correspond to the independence property $U \perp\!\!\!\perp (T, Z)$ of the graph and to the missing $Z \to Y$ arrow. With the aid of this graph, we can derive the following equivalent set of independences expressed in terms of domain variables (rather than of counterfactuals): $Y \perp\!\!\!\perp (F_Z, F_T) | (Z, T)$ and $Y \perp\!\!\!\perp Z | (U, M, T)$. The former states that the value of Y develops out of (Z, T) in a way that does not depend on how these two variables have been generated (no confounding). The latter states that Z does not directly influence Y.

We conclude by illustrating the main points with the aid of a study of the role of the acid sensing ion channel 1 (ASIC 1) in the development of multiple sclerosis. The study is a collaboration with Luisa Bernardinelli, of the University of Pavia. The treatment here consists of inducing in each experimental mouse a disorder called experimental autoimmune encephalomyelitis (EAE), that simulates the neuropathological changes of human multiple sclerosis. The mice are randomized over two levels of severity of the induced EAE, the level of severity being represented in the diagram by the binary variable T. Each mouse is also characterized by the genotype at the rs28936 locus, which regulates the expression of ASIC 1 and is represented in our diagram by the three-level variable Z, the number of copies of the deleterious rs28936 allele. Induction of EAE, and the consequent inflammatory process, causes an increase in the expression of ASIC 1, and a corresponding neurological deficit, that we record in each mouse, in the form of an ordinal score, Y, after 15 days from inoculation. Also recorded, in each mouse, is the level of ASIC 1 expression, M, in terms of the amount of messenger ribonucleic acid in neuronal cell bodies at 15 days from inoculation. Of inferential interest is the extent to which the effect of EAE (node T) on the deficit (node Y) is mediated by quantitative changes in ASIC 1 expression (node M), and by the consequent increase in ion influx. The study can be adapted to the proposed parallel encouragement design, with the genetic effect acting as encouragement. Compliers, in this example, are all mice in which presence of the deleterious rs28936 allele induces an increase in ASIC 1 expression. Knowledge of molecular mechanisms supports the claim that such compliers represent a stable majority of the mouse population. Hence, under the assumptions that are represented in our causal diagram, the method proposed by the authors can be used to calculate meaningful bounds on the $T \to M \to Y$ indirect effect, the effect that inflammation exerts on disease severity via changes in ASIC 1 expression.

It is a privilege for me to have been invited to discuss a paper which will no doubt stimulate plenty of future research.

I therefore have great pleasure in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

Guanglei Hong (*University of Chicago*)

I congratulate Kosuke Imai and his colleagues for another important methodological paper on identifying causal mechanisms. The experimental designs that they proposed have many attractive features. Yet I am concerned with the assumption of no treatment-by-mediator interaction in the parallel designs and the assumption of no carry-over effect in the crossover designs. We can find many applications in social sciences in which these two assumptions are implausible.

I propose a 'covariate-informed parallel design' that does not require these key assumptions. This new design is similar to the parallel design except that the second experiment employs covariate-informed randomization in the same spirit as a randomized block design.

Let $D=0$ and $D=1$ denote the first and second experiments respectively. For simplicity, let treatment T and mediator $M(t)$ both be binary. After collecting pretreatment information $\mathbf{X}$, we randomly assign participants to either $D=0$ or $D=1$.

Participants in the $D=0$ group are assigned at random to either $T=0$ or $T=1$. We observe $M(t)$ and specify a prediction function relating $\mathbf{X}$ to $M(t)$ for $t=0, 1$.

Those in the $D=1$ group are assigned at random to either $T=0$ or $T=1$. Applying the prediction functions that are specified in the first experiment, we obtain, for each participant assigned to treatment t in the second experiment, $\phi(t, \mathbf{X}) = \mathrm{pr}\{M(t)=1|T=t, \mathbf{X}\}$. The participants are then assigned at random to $M(t)=1$ with probability $\phi(t, \mathbf{X})$. Analogous to a two-stage adaptive design in clinical trials (Bauer and Kieser, 1999; Liu *et al.*, 2002), the covariate-informed randomization should have a higher compliance rate than a simple randomized design.

In the covariate-informed parallel design, treatment and mediator are both randomized. This design requires the stable unit treatment value assumption and the consistency assumption. If using pretreatment information to create blocks, we may estimate the block-specific treatment effects as well as the average treatment effect. By comparing each of these effects across the two parallel experiments, we may partially test the consistency assumption. In the second experiment, we may test the no treatment-by-mediator interaction assumption not only on average but also within each block.

More importantly, when the no-interaction assumption fails, researchers can nonetheless apply ratio of mediator probability weighting to estimate the counterfactual outcome $E[Y\{1, M(0)\}]$ (Hong, 2010; Hong *et al.*, 2011). For a participant assigned to $T=1$ and to mediator value m in the second experiment, the weight is

$$\omega = \frac{\mathrm{pr}\{M(0)=m|T=0, D=1, \mathbf{X}\}}{\mathrm{pr}\{M(1)=m|T=1, D=1, \mathbf{X}\}}.$$

We can show that $E(\omega Y|T=1, D=1) = E[Y\{1, M(0)\}]$. Future research may investigate the sensitivity of results to the specification of the prediction functions.

Brian L. Egleston (*Fox Chase Cancer Center, Philadelphia*)

I enjoyed this paper. The authors provide useful details on assumptions that are needed to identify mediational pathways. I do worry, however, whether we are doing scientists a disservice by focusing on indirect and direct effects as targets of investigation. Some of the interest in indirect and direct effects can probably be tied back to Wright's (1921) work on path analysis. Many scientists might be using the outgrowth of path analytic techniques without considering whether the estimands are germane to their research.

Imai, Tingley and Yamamoto have a particular focus on 'natural' effects (Pearl, 2001), as shown in equations (2) and (3) of their paper. Natural effects are not necessarily useful in cancer therapeutic development. A current goal of much research is to identify causal pathways of cancer growth that can be blocked. Although this research has led to the creation of useful drugs, the therapeutic effect has often been less than ideal. One problem is that the human body has built-in biologic redundancy. Hence, if a pathway is blocked, the body will often find another mechanism to achieve the same goal. This has led to estimands of interest that differ from those of focus by the authors.

Notationally, let G_z represent cancer-related gene number z for $z=1, \ldots, n$ ($G_z=1$ if G_z is active and $G_z=0$ otherwise). Let $T(C)$ represent survival time under cancer state C ($C=0$ if no cancer and $C=1$ if cancer). Let $T(C, G_1)$ and $T(C, G_1, G_2)$ represent potential survival outcomes under G_1 alone and with G_2. Current therapeutic research is interested in creating a situation in which $E[T(1)] = E[T(0)]$. Using inhibitors, G_z becomes manipulable. A first step in development is to investigate whether $E[T(1, 0)] = E[T(0)]$. Unfortunately, scientists generally find that $E[T(1, 0)] < E[T(0)]$. However, in the course of investigating why the survival benefit when inhibiting G_1 is not as great as expected, researchers discover that G_2 has taken

over many of the functions of G_1. Previously, G_2 was not strongly implicated as a potential confounder or mediator. A new inhibitor of G_2 is developed and investigators find that $E[T(1,0)] < E[T(1,0,0)] < E[T(0)]$, and the cycle of discovering why blocking pathways is not as successful as intended continues.

Although G_j might be a mediator of the relationship of $T(C)$ and G_{j-1} for $j > 1$, the relationship is not necessarily discoverable until G_{j-1} is inhibited. The estimation of $E[T(1,0)]$ and $E[T(1,0,0)]$ involves manipulation of the mediators, and the natural effects are of little inherent interest.

Roland R. Ramsahai (*University of Cambridge*)

The paper computes bounds on $\bar{\delta}(t)$ for the simple experiment design, assuming that Y_i is deterministically related to (M_i, T_i). This is computed from $\bar{\tau} = \bar{\delta}(t) + \bar{\zeta}(1-t)$ and the bounds on $\bar{\zeta}(1-t)$ in Sjölander (2009). For the decision theoretic direct and indirect effects, $\bar{\tau} = \bar{\delta}^{\dagger}(t) + \bar{\zeta}^{\dagger}(1-t)$ (Didelez *et al.*, 2006; Geneletti, 2007) and the bounds on $\bar{\zeta}^{\dagger}(1-t)$ and $\bar{\zeta}(1-t)$ are identical (Ramsahai, 2012). Therefore these bounds are valid within the decision theoretic framework (Dawid, 2002), which involves no determinism.

The paper also derives bounds on $\bar{\delta}(t)$ for the parallel experiment design, assuming that Y_i is deterministically related to (M_i, T_i). Let $\delta_U^{\dagger}(t)$ be the individual indirect effect, where U represents the relevant individual characteristics. Since $U \perp\!\!\!\perp T | D = 0$ and $U \perp\!\!\!\perp (T, M) | D = 1$

$$\left.\begin{array}{ll} P_{ymt} = E_U(p^U_{ym|t}), & p^U_{ym|t} = \theta^U_{ymt}\phi^U_{mt}, \\ Q_{ymt} = E_U(p^U_{y|mt}), & p^U_{y|mt} = \theta^U_{ymt}, \\ \bar{\delta}^{\dagger}(t) = E_U\{\delta^{\dagger}_U(t)\}, & \delta^{\dagger}_U(t) = (\theta^U_{11t} - \theta^U_{10t})(\phi^U_{11} - \phi^U_{10}), \end{array}\right\} \tag{11}$$

where $p^U_{ym|t} = \mathbb{P}(Y = y, M = m | T = t, U)$, $\theta^U_{ymt} = p^U_{y|mt}$ and $\phi^U_{mt} = p^U_{m|t}$. From expression (11), the method of Dawid (2003) and Ramsahai (2007) obtains identical bounds on $\bar{\delta}^{\dagger}(t)$, as $\bar{\delta}(t)$, in terms of (P_{ymt}, Q_{ymt}). Thus the bounds in the paper are applicable without determinism.

Let $\sigma_T \in \{t, \emptyset\}$ and $\sigma_M \in \{m, r_t, \emptyset\}$ represent the strategies for assigning the values of T and M, where $\sigma_M = \emptyset$ represents observation, $\sigma_M = m$ represents that M is assigned a value m by randomization and $\mathbb{P}(M | T, U, \sigma_M = r_{t^*}) = \mathbb{P}(M | U, \sigma_T = t^*)$. It can be shown that $\bar{\delta}^{\dagger}(t)$ is identifiable with the expression in theorem 1 if

$$U \perp\!\!\!\perp (\sigma_M, \sigma_T), \tag{12}$$

$$Y \perp\!\!\!\perp \sigma_M | M, U, \sigma_T = t, \tag{13}$$

$$M \perp\!\!\!\perp \sigma_T | T, U, \sigma_M, \tag{14}$$

$$\mathbb{P}(Y | U, \sigma_M = m, \sigma_T = t) - \mathbb{P}(Y | U, \sigma_M = m, \sigma_T = t') = g(t, t', U). \tag{15}$$

The potential outcomes probabilities are invariant to the value assigned by randomization, by definition, and the paper assumes that they are invariant under randomization or observation. This invariance is no weaker than condition (12), which restricts the distribution of the individual characteristics U to be invariant to the strategy for assigning T and M.

The notation $Y_i(t, m)$ in the paper assumes that, given $T = t$ and $M = m$, the strategy for obtaining these values is irrelevant (Cole and Frangakis, 2009). This notation is justified from assumption 3 and the standard implicit consistency assumptions $Y_i(t) = Y_i\{t, M_i(t)\}$ and $Y_i\{t, M_i(t'), d\} = Y_i(t, m, d)$ if $M_i(t') = m$. Such assumptions are as strong as condition (13). The paper assumes further consistency by $M_i = M_i(T_i)$, i.e., for an individual, the value of M when $T = t$ is observed is the value of M when $T = t$ is assigned by randomization. This is no weaker than condition (14). Since condition (15) is a no-interaction assumption, the conditions for identifying $\bar{\delta}(t)$ in the parallel design are as strong as those for identifying $\bar{\delta}^{\dagger}(t)$. Similar comments apply to other experimental designs.

Chen *et al.* (2007) showed that, under a zero direct effect, the causal effect of T on Y is not predictable from the effect of T on M and M on Y. Conditions were given in Chen *et al.* (2007) to ensure that the effect is predictable from the chain of effects. Perhaps similar criteria can be developed under a non-zero direct effect and then used to develop tests to check the validity of the 'causal chain' approach in Section 3.

Theis Lange (*University of Copenhagen*)

Firstly I congratulate the authors for an important and enjoyable paper; secondly I thank Professor Imai for an inspiring presentation at the Society. On reading the paper I was left with two concerns or perhaps more accurately wishes for future research.

(a) In the parallel design we have replaced the (untestable) assumption of sequential ignorability with an assumption of no-interaction at the unit level (which at least has testable implications). However, for non-binary outcomes the no-interaction assumption is scale dependent. I fear that it would often be difficult to argue for the validity of the no-interaction assumption by using only subject matter knowledge, even when there are good subject matter arguments for a mechanistic causal effect separation since such arguments are rarely scale specific. Thus, we have replaced a structural assumption (namely sequential ignorability) with a purely technical assumption. Perhaps future research can either remove this assumption or establish whether we are still estimating something interesting when the no-interaction assumption fails.
(b) On the basis of the present paper it could be argued that for any new experiment aiming at quantifying causal mechanisms one of the novel designs should (if at all possible) be employed simply as a precautionary measure. However, before adopting this guideline it would be of great value to know the price we are paying in terms of statistical uncertainty. Or, in other words, assuming that both sequential ignorability and the no-interaction assumption hold, but we only have 100 study subjects, are these 100 subjects then best 'used' in a single-experiment set-up or a parallel design in terms of statistical uncertainty of the resulting estimators?

Andrew Gelman (*Columbia University, New York*)

This is an impressive paper that goes beyond philosophical argument and mathematical manipulation and proposes specific designs to study real problems. Several of the proposed new studies seem fairly inexpensive—e.g. the expanded survey experiment in Section 4.2.3 on attitudes towards immigration—and I wonder whether the authors are considering performing these studies themselves or perhaps know of others who have such plans. Often in political science and economics we need to wait for new data (new elections; new revolutions; new economic or political trends), but these psychological studies can be replicated fairly easily, and I am curious about the results.

I have two further questions, one applied and one methodological. My applied question is about the effect of incumbency and money in US congressional elections. Unlike many causal questions in social science, this one can be formulated cleanly: for incumbency, either an incumbent runs for re-election or not. For money, if I give \$100 to candidate X, what is the expected effect on his or her vote share in the upcoming election? Also, whether an incumbent runs for re-election affects the campaign contributions in his or her district. Although all these effects are clearly defined, studying them is tricky: incumbents' decisions, results of primary elections and campaign donations are observational variables, as are the aspects of their opponents in the general elections. There is a literature on the estimation of the effects of incumbency and money on elections using various clever ideas with observational data and natural experiments. I am wondering whether the methods described in the present paper can be applied in this observational setting.

Finally, just as a minor comment: I hope in the future that the authors will think as hard about the presentation of their results as they do about their mathematical foundations. For example, they estimate a proportion as '0.730 for the treatment group and 0.392 for the control group'. Given that their sample sizes are below 68 and 198 respectively, I think that third digit is meaningless. Similarly, they present a confidence interval as $[-0.175, 0.825]$. Given the evident level of uncertainty, $[-0.2, 0.8]$ would suffice. One of the most important messages statisticians convey is about the presence of uncertainty, and we dilute much of this when we display meaningless levels of precision.

Manabu Kuroki (*Institute of Statistical Mathematics, Tokyo*)

I congratulate the authors on this paper which tackles a difficult but interesting problem. I would like to provide some comments on the present paper from the viewpoint of quality control (QC) which is one of my main research fields.

Dr Genichi Taguchi, who was a pioneer of quality engineering, implied that the effect decomposition problem is one important issue in experimental design (Taguchi (l987), chapter 28). However, he did not provide any solution to this problem and this problem has not attracted much attention from QC experimenters for a long time. In this sense, although the authors' research area is different from that of Dr Taguchi, the authors also shed light on the effect decomposition problem in experimental studies. Thus, the present paper provides a new motivation for QC practitioners who deal with this problem.

The results of the present paper have some limitations when we apply them to the QC area:

(a) the assumption of no carry-over effects does not hold in many cases;
(b) the monotonicity assumption is often violated and

(c) a main interest in QC is to evaluate direct and indirect effects in the whole population instead of a subset of the population.

Despite these limitations, the present results may be applicable to, for example, an experimental study to evaluate the effects in the case where we exchange components that may cause bad performance and the deterioration of assemblies (or sequential systems). To overcome these limitations, as one solution, the authors provided sharp bounds in some cases but they did not formulate bounds for encouragement designs. It would be helpful if the authors can provide the formulation of sharp bounds for these designs because sharp bounds often provide useful information on the evaluation of direct and indirect effects (e.g. Cai *et al*. (2008)).

The following contributions were received in writing after the meeting.

Jeffrey M. Albert (*Case Western Reserve University, Cleveland*)
I commend the authors on a stimulating and clearly written paper, and one that is a welcome contribution to the limited literature on study designs for the assessment of mediation. For brevity, and to cover the main ideas, my comments are focused on the (two-part) parallel design (with direct or imperfect manipulation).

The two-part design has some appealing features. In particular, it clearly separates the goal of estimating the overall treatment effect (which is provided by the first experiment) and that of estimating direct and indirect effects (which are provided by the second experiment). However, because the second experiment does not contribute to the estimation of the overall treatment effect (except possibly with additional assumptions), an obvious drawback of the design proposed is the requirement of additional resources for the estimation of mediation effects. It may be argued that researchers, or funding agencies, should be willing to pay the price for this information. However, when resources are limited this may be a difficult sell. In contrast, in the standard 'single-experiment' design, for which the primary objective is usually inference for the overall treatment effect, mediation analysis is offered 'for the same price', albeit with additional strong assumptions. Of course, low power for testing mediation effects may require a boost in the sample size, but then inference for the overall treatment effect will also benefit. To allow a complete evaluation, the power implications of the proposed *versus* standard designs could use further investigation.

It is notable that, even with the additional investment that is represented by the two-part design, the estimation of mediation effects still requires strong assumptions that are not assured by the randomization. These assumptions include that of no (individual level) treatment–mediator interactions and the 'consistency assumption' (assumptions 3 and 9). It is interesting that assumptions 3 and 4 (or 8 and 9) essentially render the Z as an instrumental variable. The authors dismiss the instrumental variable approach; however, some generalized (e.g. two-stage least squares, extending Albert (2008)) approach may be possible without having to assume no direct effect of T on Y (noting that multiple instrumental variables may be obtained from the multiple-category Z). Unfortunately, the assumption of no direct effect of Z (as well as T) on Y may be implausible in many situations, in which cases it is not clear whether it is worth trading this assumption for that of sequential ignorability.

John G. Bullock (*Yale University, New Haven*) **and Donald P. Green** (*Columbia University, New York*)
Imai, Tingley and Yamamoto remind us that an intervention's 'direct effect' and 'indirect effect' are fundamentally unidentified. Both involve inherently unobservable potential outcomes. Not even a randomized experiment can render an estimate of quantities such as $E[Y_i\{t, M_i(1-t)\}]$. Yet they express optimism about our ability to learn about direct and indirect effects by coupling experiments with an array of supplementary assumptions. We applaud them for detailing the assumptions that are required to isolate causal mechanisms. But, when reflecting on applications in the social sciences, we remain sceptical about whether any experimental design will permit a researcher to estimate direct or indirect effects convincingly.

We are sceptical because the assumptions that are invoked by the authors are not directly testable: the 'consistency assumption' (assumption 3), the 'no-interaction' assumption (assumption 5) and the homogeneous unit effects assumption (on the eighth page). In practice, the list of assumptions in social science applications is even longer. First, social scientists routinely study mediation by using variables, such as beliefs or feelings, that are not observed directly. It is difficult to measure and manipulate a particular mediator without inadvertently measuring and manipulating other mediators as well. Measurement challenges are especially daunting given widespread reliance on survey measures; subjects are often invited to report beliefs or feelings, and their responses are used to measure the mediator and the outcome. Systematic response error that affects both mediator and outcome is a very real possibility.

Second, rarely can social scientists set specific values of a mediator (i.e. $M(t)$). At best, they intervene by using 'encouragement' designs like those which the authors discuss in Section 4. These designs force researchers to invoke additional untestable assumptions: most notably, the 'exclusion restriction', which says that encouragements affect outcomes solely through the intended mediator.

Few, if any, social science studies have satisfied or could convincingly satisfy these assumptions. Rather than attempt to estimate parameters that are fundamentally unidentified, let us set our sights on the still challenging task of estimating the causal effects of t and M. Even if we cannot know the indirect effect of t, we can still learn about its effects on hypothesized mediators, and we can learn the average effect of intervention-induced change in M on outcomes. The advantage of this approach is that it puts us on a firm experimental footing. After we have accumulated substantial knowledge about the effects of t and M, identification of causal mechanisms may become more plausible.

Vanessa Didelez (*University of Bristol*)

The importance of experimental design for causal inference is twofold. It can guarantee crucial assumptions; for example actual randomization allows identification of average causal effects. A careful design also clarifies, almost defines, the target of inference—this is especially relevant in the context of sometimes woolly notions of 'causal mechanisms'.

The authors consider indirect or direct effects involving $Y\{t, M(t')\}$. Setting treatment to t and t' for the same unit is genuinely counterfactual Consequently, although their designs improve on the single experiment, they cannot avoid untestable assumptions. Furthermore, do the designs proposed *define* the target of inference? The additional experiment in the parallel designs really targets the controlled direct effect, and the two require linking by untestable assumption 5. However, the crossover designs proposed clearly target $Y\{t, M(t')\}$, and under untestable assumption 7 this comes close to observing $Y\{t, M(t')\}$ itself.

A different type of design is sometimes possible and clarifies the causal parameter in a decision theoretic context (Didelez *et al.*, 2006), namely when we can manipulate the mediator, without controlling it (almost) as if treatment were at two different values for the same unit. For example double-blind placebo-controlled studies: these target the direct effect of an active ingredient not mediated by the patient's or doctor's expectation. Crucially the mediator (the expectations) is not (and cannot) itself be controlled, but the design guarantees that it arises as under 'drug taken'. One can easily think of variations addressing the indirect effect, here the placebo effect. This type of design seems feasible whenever 'treatment' comprises different aspects that could—with a little imagination—be separated out. Robins and Richardson (2010) used similar examples and (possibly hypothetical) interventions in augmented directed acyclic graphs to discuss when $Y\{t, M(t')\}$ can be regarded as a manipulable quantity. Does this mean that we observe $Y\{t, M(t')\}$ itself? Not necessarily: the design fails when there are post-treatment confounders (even if observed) of M and Y; in the placebo-controlled trial this is known as 'unblinding', e.g. by side effects of the active ingredient.

Looking at typical applications, it will be rare that crossover or placebo-type designs can be used. The interest in causal parameters based on $Y\{t, M(t')\}$ therefore remains a mystery to me—what *practical* questions does it help to answer that simpler approaches (causal chain or controlled effects) do not? If effect modification is the main problem, we should maybe direct more attention to investigating effect modification and design experiments accordingly.

David Draper (*University of California, Santa Cruz*)

The authors of this interesting paper have offered us some increased clarity on a difficult question: can we go beyond estimating the average effects of causes to correct identification of the actual underlying causal *mechanisms*? Their answer is a cautious yes, by employing designs they recommend that differ from those in widespread current use; I am less sanguine, for at least the following two reasons.

(a) It is distressingly easy to imagine experiments in which the authors' assumption 3, which they correctly point out is crucial to their attempts at improved designs, does not hold. For example, consider an experiment in which the dichotomous treatment variable is a form of talk therapy aimed at behaviour modification to avoid out-of-wedlock pregnancy ($T = 1$) or no such therapy ($T = 0$), and the outcome variable is the number of sexual partners. To keep this example from being too stereotypically gendered, imagine a world in which an effective male contraceptive pill is available, and consider one of the authors' designs in which use or non-use of this pill is the mediator to be manipulated, on a cohort of young men. It is a brave (and foolhardy) assumption in this setting to believe that a man who chooses to take the pill will behave identically to a man who is randomized to the pill with respect to the number of sexual partners that he seeks.

(b) Almost all the authors' examples involve a single mediator, but what if (as will often be the case) two or more mediators are active (i.e. highly relevant to correct causal conclusions, because of strong correlations with the treatment and outcome variables) but you are aware of only one of them, and therefore—using one of the authors' designs—you manipulate only the one that you know about? Then what looks to you like unexplained variability in the outcome may actually be bias (arising from having manipulated only one mediator), and this will potentially distort your causal conclusions.

A little more detail on the following point would also be helpful. The authors make frequent use of the expectation operator, without saying what distribution the expectation is over: are we averaging over the distribution yielding the randomization to experimental groups distribution (holding experimental subjects constant, and attempting to generalize only to what the results would have been if they had ended up in different groups), or the distribution that is implied by the usual (often unstated, and often untrue) assumption that the subjects are like a random sample from the population to which we are actually trying to generalize, or what?

Adam N. Glynn (*Harvard University, Cambridge*)
This paper provides a thorough investigation of potential solutions to a difficult problem. As the authors note, much of this difficulty stems from the fact that, although designs with direct or indirect manipulation of the mediator provide more information about the mediation effect, these designs also require that the manipulation does not directly affect the outcome.

Interestingly, by clarifying these difficulties, this paper may lead researchers in non-experimental settings to reconsider whether mediation is the question that they want to address. For example, in observational studies of racial discrimination, the treatment could be conceptualized as the perception of race at the time of application (instead of race defined at birth as in the example from Section 3.2.4). This allows an applicant's qualifications to be incorporated in the analysis as pretreatment variables, and mediation analysis would not be necessary (see Greiner and Rubin (2011) for a discussion).

As another example, consider the conjecture known as the weak states mechanism—that natural resource abundance (e.g. oil or diamonds) might reduce the incentive for a state to develop the bureaucratic capacity that is necessary for taxation, and that this lack of state capacity might increase the likelihood of civil conflict (Fearon and Laitin, 2003). One reason why we might want to study this mechanism is to anticipate the effect of laws that would block the mediation effect (for example see the discussion of oil revenue management laws in Humphreys (2005)). However, any such intervention might have its own direct effects on the outcome and, therefore, the mediation effect may not necessarily represent the effect of interest.

It is unclear to me whether this paper will do more to encourage the use of good design or to dissuade questionable (and possibly unnecessary) attempts at mediation analysis. In either case, the authors have done a great service in clarifying the issues.

Booil Jo (*Stanford University*)
I congratulate the authors on their very important and stimulating contribution to the causal inference literature. Possibilities of manipulating mediators have been largely overlooked and, therefore, little knowledge has been accumulated so far about design possibilities in identifying causal mechanisms. It may seem that the proposed alternative experimental designs replace one untestable assumption with another set of untestable assumptions (that could be even stronger). However, these alternative experimental designs let us explore alternative identifying assumptions, the use of which is likely to improve the quality of our causal inference. As the authors emphasized, when the single-experiment design is the only option, the unavoidable choice of identifying assumption is sequential ignorability, which is not a desirable situation. The use of alternative designs and identifying assumptions opens up possibilities for diverse and improved sensitivity analysis strategies. Further, the authors demonstrated the use of encouragement, which not only makes implementation of the designs suggested more feasible but also improves the testability of some of the underlying identifying assumptions.

What seems somewhat unclear at this point is how the design strategies suggested will pan out in practice. The designs proposed will generally require larger sample sizes. This may not be feasible in many studies that must rely on small to moderate sample sizes. For example, in many medical and health-related experiments, recruiting a large sample is simply not feasible. The suggested parallel designs consist of two experiments, which inevitably require larger sample sizes. Even if recruitment is possible, the increased cost and practical issues that are related to having two experiments may discourage the use of the designs suggested. The crossover designs seem less costly, but the no-carry-over effects and consistency assumption

can be quite strong. To make this assumption more testable, a larger sample is again needed to maintain the same level of statistical power (i.e. we need to include a group of individuals without mediator manipulation). I also suspect that we shall need some guidelines on ethical issues related to manipulation of mediators. Finally, I wonder how applicable the study designs suggested are. The examples that are used in the paper (transcranial magnetic stimulation and immigration) seem quite unique, making me somewhat unsure about the broad use of the designs suggested. I look forward to seeing more applications in diverse settings. I congratulate the authors again and hope that that this paper will ignite further development of creative and practical study designs to elucidate causal mechanisms.

David A. Kenny (*University of Connecticut, Storrs*)
The paper is in the now rather old tradition of finding ways of estimating causal mechanisms by combining experimental and non-experimental approaches. I have three comments.

First, the authors' approach is to estimate the indirect effect (IE) as the difference between the total effect of T on Y and the direct effect of T on Y controlling for M. Such an approach is implicit in Baron and Kenny (1986) and was formally described in Clogg *et al.* (1992). An alternative, less general, but currently quite widely utilized, strategy for the estimation of IEs is to estimate the IE as the product of two effects: the path from T to M or a and the path from M to Y or b. Where appropriate, knowing the sizes of a and b can be very informative. First, if the IE is near 0, it is useful to know whether path a or b (or both) is 0. For instance, if path a is 0 but b is not, then we know that the intervention failed to trigger the mediator. Second, the relative size of path a and b can be informative. Some mediators are 'proximal' in that they are closer to T (Hoyle and Kenny, 1999) whereas others are 'distal' in that they are closer to Y.

Second, I think it highly unlikely that one ever has a 'pure' manipulation of M and so the authors' consideration of such seems misplaced. In the tradition of Cook and Campbell (1979), a measure or manipulation is virtually never identical to the construct that it purports to measure. Moreover, mediators are typically inside the 'black box', and so they can be difficult to observe directly. It should also be realized that almost always the manipulation of T is one of 'encouragement', and so the use of encouragement is not a poor second choice. Rather it is what is almost always done.

Third, in cases for which we can assume continuous M and Y, no TM interactions and linear effects, I think that a single experiment can be undertaken in which both T and M are manipulated and measured. In such situations, the IE could be measured as the product of two effects, ab. The single experiment would yield a more precise estimate of the IE than the two-arm study proposed by the authors. The interested reader can consult Smith (1982) for an instructive example.

Víctor Leiva and Emilio Porcu (*Universidad de Valparaíso*)
This interesting paper deals with designs of randomized experiments to evaluate the treatment effect on a response under causality, where the treatment effect is the sum of the causal mediation indirect (mediator) and direct effects. Although the single design is one of the most commonly used methods for identifying causality, it is based on assumptions that are difficult to justify in practice. The paper proposes parallel and crossover experimental designs by means of which it is possible to manipulate the mediator that connects the treatment and response. These designs are based on a key assumption that is the consistency, which allows us to manipulate the mediator without directly affecting the response. These designs improve the results from the single design.

Studies in diverse areas are usually causal and not associational. This makes standard statistical inference not suitable for these studies and so-called causal inference is needed instead. In general, because studies in these areas are usually observational and not experimental, it is somewhat complicated to justify parametric assumptions and so the use of semiparametric models seems to be more adequate. Indeed, there are examples where to assume parametric models implicitly leads to models that exclude *a priori* the null hypothesis of no causal effects; see Robins and Wasserman (1997). In spite or these difficulties, some efforts on the use of parametric models in causal inference, including non-normal distributions, have been made; see Shimizu and Kano (2008).

In parametric modelling, it is well known that outliers produce undesirable effects on the estimates of the model parameters, influencing their behaviour. Then, it is important to have tools that allow us to assess such influence. A method known as local influence provides us with an instrument to detect the effect of small perturbations in the model on the parameter estimates; see, for example, Leiva *et al.* (2007) and references therein. Because the problem of influence could also be present in causal models, with similar consequences, the idea of influence diagnostics could be explored in the class of models analysed in the paper.

Outcomes, mediators (such as 'anxiety') and direct effects can be accumulated in a similar way to that generated by a fatigue process, which acts under stress. Then, the data-generating process could be

well explained by a process of this kind and so a non-normal model, such as the Birnbaum–Saunders distribution, might be considered in causal analyses of the type studied in the paper; see Leiva *et al.* (2007).

N. T. Longford (*SNTL and Universitat Pompeu Fabra, Barcelona*)

Statistical literature is replete with poorly founded claims of having identified causes and generated some understanding of causal mechanisms. This paper is a commendable effort to add scientific rigour to the discourse about causal mechanisms and to the design for studying them. However, the framework presented is not particularly constructive, because the numerous assumptions, although well motivated, are presented in the form of imperatives—if a particular setting departs from a required assumption, the edifice that is essential for the inference crumbles. The fact that some assumptions are unverifiable, or even untestable, adds to the difficulties. A more constructive approach would define metrics for departures from the assumptions and allow for some form of arbitration about how great a deviation from the assumption (the ideal) is permitted without undermining the inference about the causal mechanism. For example, carry-over in a (clinical) crossover trial can rarely be regarded as absent (satisfying the relevant null hypothesis H_0), because such an absence corresponds to an unsupportable H_0. Failure to reject H_0 does not suffice here, even if we have ample evidence from elsewhere that the carry-over is sufficiently small for a different purpose.

I think that the limitation of the presented methodology to very simple causal mechanisms is not made clear. A unit (or link) of a causal mechanism is a direct cause without a mediator. All the examples discussed are mechanisms comprising two units. In more realistic settings, there are many interrelated mediators, and the framework presented would entail a large set of interrelated experiments and randomizations. For example, in the study of attitudes to immigration, having been abroad, having contemplated living there, having acquaintances among immigrants, having an occupation that involves international contacts, and the influence of the (self-selected) media outlets are relevant factors, most of them beyond our ingenuity and resources to manipulate. To study a causal mechanism (the verb 'to identify' is misleading because it implies a verdict with certainty that cannot be arrived at by a hypothesis test on a finite sample), we must have the ability to manipulate each mediator in a way that is described by the assumptions (extended to settings with several mediators), and that is a rather tall order.

David P. MacKinnon (*Arizona State University, Tempe*)

Imai, Tingley and Yamamoto link experimental designs and modern causal inference, thereby clarifying limitations about what experiments can demonstrate regarding a mediating mechanism. This important work is applicable to the many areas where researchers seek understanding of how a manipulation affects an outcome. I do not agree that the single-experiment design is how mediating mechanisms are identified. The search for mediating mechanisms is addressed by a programme of experimental research, replication studies, history and qualitative data, conducted by different researchers in different research contexts (MacKinnon, 2008). It is unlikely that any one study, even the ideal experiment designs that are described in the paper, would be sufficient to identify a mediating process (because of type II errors, for example). A programme of research is also critical to deal with other considerations, such as the requirement of valid and reliable measures, sample representation of the population of interest and selection of the position in a chain of mediation to investigate.

Given the strong assumptions that are necessary for identifying mediating mechanisms, it would seem surprising that mediating mechanisms can be found. However, research that is focused on predicted and observed patterns of results in different contexts is how mediating processes have been identified in the past. A few notable mediating mechanisms are atomic theory in chemistry, gene theory in genetics and cognitive dissonance theory in social psychology. In the social sciences, several designs that are closely related to those in the paper have been used to test logical predictions of mediation theory (Mark, 1986; MacKinnon 2008; MacKinnon and Pirlott, 2010). In the social science literature, the parallel and encouragement designs correspond to blockage and enhancement designs where additional conditions are specified that should lead to larger or smaller effects on outcomes depending on whether the mediator was enhanced or blocked. Also related are double-randomization designs whereby a manipulation is conducted and a mediator and outcome measured, and then a second randomization addresses the mediator-to-outcome link. Other designs attempt to demonstrate specificity for a mediation process by predicting mediation through a hypothesized mediator and not through a comparison mediator. Useful future research would clarify the causal assumptions of these additional designs, including methods to address the sensitivity of conclusions to assumptions. Another valuable next step is the application of experimental designs to

answer important substantive questions with real data that includes collaboration between substantive researchers and statisticians.

Jorge Mateu (*University Jaume I, Castellón*), **Oscar O. Melo** (*National University of Colombia, Bogotá*) **and Carlos E. Melo** (*District University Francisco José be Caldas, Bogotá*)

Identifying causes is the goal of most scientific research. We can design research to create conditions that are very comparable so that we can isolate the effect of the treatment on the dependent variable. In this way, research designs that allow us to establish these criteria require careful planning, implementation and analysis. Many times, researchers must leave one or more of the criteria unmet and are left with some important doubts about the validity of their causal conclusions, or they may even avoid making any causal assertions.

We would like to draw the authors' attention to a particular problem that could benefit from this strategy. To improve further on the crossover design, the results can be extended to models with the observed pretreatment covariates X_i. Then, the average indirect effect by using the same notation as the authors' is given by

$$\bar{\delta}(t) = \mathbb{E}[Y_{i1}\{t, M_i(1)|X_i = x\}] - \mathbb{E}[Y_{i1}\{t, M_i(0)|X_i = x\}]$$

for $t = 0, 1$ and all $x \in \chi$, and where $M_i \in \mathcal{M}$ denotes the observed value of the mediator that is realized after the exposure to the treatment, $\mathcal{M}$ is the support of M_i and the two potential values $M_i(0)$ and $M_i(1)$ are the effects of the treatment over the mediator. During the second period of the experiment, the treatment status is $1 - T_i$ for each unit, and the value of the mediator equals the observed mediator value from the first period, M_i. So, in the second period, the observed outcome can be written as $Y_{i2} = Y_{i2}(1 - T_i, M_i|X_i)$. The following assumption is satisfied under the crossover design because the treatment is randomized:

$$\{Y_{i1}(t, m), Y_{i2}(t', m), M_{i1}(t'') : t', t'' \in (0, 1), m \in \mathcal{M}\} \perp\!\!\!\perp T_i|X_i = x$$

for $t, t' = 0, 1$. Additionally, Robins (2003) and Imai *et al.* (2010) considered the identification. In this case, it should satisfy the following assumptions:

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i|X_i = x \text{ and } Y_i(t', m) \perp\!\!\!\perp M_i(t)|T_i = t, X_i = x,$$

where it is also assumed that $0 < P(T_i = t|X_i = x)$ and $0 < P\{M_i(t) = m|T_i = t, X_i = x\}$ for $t = 0, 1$ and all $x \in \chi$ and $m \in \mathcal{M}$. To this consistency assumption, the absence of carry-over effects can be also assumed, i.e.

$$\mathbb{E}[Y_{i1}\{t, M_i(t)|X_i\}] = \mathbb{E}[Y_{i2}(t, m|X_i)] \qquad \text{if } M_i(t) = m$$

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

Alessandra Mattei (*University of Florence*)

Imai, Tingley and Yamamoto provide a valuable contribution on a subject that is just as attractive as it is challenging: understanding causal mechanisms. They focus on natural direct and indirect effects, which are defined as a function of potential outcomes of the type $Y_i\{t, M_i(t')\}$, $t' \neq t$, usually named '*a priori* counterfactuals', because they cannot be observed for any subset of units in a specific experiment.

To embed natural direct and indirect effects in the potential outcomes framework formally, the primitive concepts and the basic assumptions for causal inference should be generalized to make potential outcomes of the form $Y_i\{t, M_i(t')\}$, $t' \neq t$, well-defined objects. Specifically, natural direct and indirect effects require that the intermediate variable M could be, at least in principle, regarded as an additional treatment. Therefore, assumptions on the compound assignment mechanism for the multivariate treatment variable (T, M) should be contemplated.

The parallel and crossover (encouragement) designs that are proposed by the authors imply that (partial) interventions on the intermediate variable are conceivable. My feeling is that, if we are willing to entertain hypothetical interventions on the intermediate variable, it could be more reasonable to design a single experiment posing a compound assignment mechanism for the treatment variable and the mediating or encouragement variable: alternative causal paths could be investigated, and various hypotheses on the causal mechanism could be assessed.

Another crucial issue concerns the assumptions of consistency and no carry-over effects, which allow us to carry out extrapolation of *a priori* counterfactuals for units on which the data contain no or little information, using data either across units for the same time or across time from the same unit. As the authors also recognize, these assumptions may be controversial: the experiment to which a unit is assigned

may make a difference, and also time may matter, implying that treatment comparisons across time lack causal interpretation.

According to me, to understand clearly the nature of alternative identifying assumptions and to obtain useful insights on how to design experiments aiming at disclosing causal mechanisms, preliminary analyses based on the principal stratification framework could be valuable. A principal stratification analysis naturally provides information on the extent to which a causal effect of the treatment on the primary outcome occurs together with a causal effect of the treatment on the intermediate outcome, without involving *a priori* counterfactuals and identification and estimation strategies based on extrapolation methods.

Emilio Porcu and Víctor Leiva (*Universidad de Valparaíso*)

The paper deals with parallel and crossover designs as an alternative to the single design, which are useful when the mediator that connects the treatment and outcome may be manipulated. The difference between these two designs is that experimental units are assigned to one of two treatments at random (parallel) or sequentially assigned to two treatments (crossover) by using the manipulation of the causal mediator. These experimental designs are based on the consistency assumption, which supposes that the manipulation of the mediator does not directly affect the outcome. By means of an example analysed in the paper, the effect of media framing on the subjects' immigration preference is tested, using the anxiety as mediator. Because the manipulation of the anxiety is imperfect, the parallel design is used, turning out to be more informative than the single design.

T. S. Richardson (*University of Washington, Seattle*) **and J. M. Robins** (*Harvard School of Public Health, Boston*)

This is a thought-provoking paper that proposes several new approaches to probing mediation. It is an attractive feature of the authors' designs that their analyses are based on counterfactual independences that hold as a consequence of randomization.

In the context of a single-intervention study where T alone is randomized, in several references, the following independence assumptions have been entertained on the basis of substantive hypotheses:

$$T \perp\!\!\!\perp Y(t, m), M(t), \tag{16}$$

and

$$M(t) \perp\!\!\!\perp Y(t, m) | T = t, \tag{17}$$

for all $t, m \in \{0, 1\}$. We have computed bounds on the average pure (or natural) direct effect (here $\mathbb{E}[\zeta_i(0)]$; see expression (3) in the main text) under these assumptions (Robins and Richardson (2011), appendix C). In the above expressions we have implicitly assumed there is a particular well-defined joint intervention that sets M to m and T to t.

Note that expression (16) follows from the assumption that T was randomized. By contrast, assumption (17) will hold in contexts in which there is no confounding between M and Y.

In situations in which it is possible to carry out the aforementioned joint intervention, we may verify assumption (17) by conducting a subsequent study in which both T and M are randomized, in the manner of the parallel design proposed by the authors. In this setting, there is a consistent test of assumption (17), i.e. it is, in principle, verifiable. Specifically, we may contrast the conditional distributions $P\{Y_i = y | M_i = m, T_i = t, D_i = 0)$ and $P(Y_i = y | M_i = m, T_i = t, D_i = 1)$ that result from the two experiments. When these distributions agree assumption (17) holds in the study where T alone is randomized (i.e. conditional on $D_i = 0$). In this case the bounds (see expressions (8) and (9) in the main text) obtained by the authors for $\bar{\delta}(1)$ imply bounds on $\mathbb{E}[\zeta_i(0)]$ that agree with those that we have obtained. As stressed by the authors, in the absence of the consistency assumption 3, the second experiment provides no information concerning the potential outcomes in the first experiment.

In contrast, no consistent test exists for the 'cross-world' counterfactual independence:

$$M(t) \perp\!\!\!\perp Y(t', m) | T = t, \tag{18}$$

even if we are willing to make assumption 3 and can carry out the parallel design. Note that expression (18) is required to obtain point identification of $\mathbb{E}[\zeta_i(0)]$ via Pearl's mediation formula.

More generally, Robins (1986) and Robins and Richardson (2011) gave a general framework for formulating causal models under which all counterfactual independence restrictions are in principle subject to experimental verification in the way that is outlined here.

Donald B. Rubin (*Harvard University, Cambridge*)

lmai, Tingley and Yamamoto are to be congratulated for addressing the challenging issue of direct or indirect causal effects using potential outcomes, a notation that was introduced by Neyman in 1923 (see Neyman (1990)) for repeated sampling, randomization-based inference in randomized experiments, and extended in Rubin (1974, 1975, 1977, 1978) to include general assignment mechanisms for treatments and other forms of inference. The condition for the notation's adequacy (e.g. discussed in Rubin (1978), pages 37–38) was eventually called the 'stable unit treatment value assumption' (Rubin (1980), page 591)—meaning that, no matter how the ith unit, $i = 1, \ldots, N$, was exposed to treatment level t, $t = 1, \ldots, T$, the outcome $Y_i(t)$ would be realized, where this could be a probability distribution (Rubin (2010), page 40); potential outcomes are functions of units and treatments at defined times of assignment of treatments and measurement of outcomes.

One component of the stable unit treatment value assumption is 'no interference'—explicit in this paper, but only implicit is the second component, 'no hidden versions of treatments' meaning that there are no levels other than those reflected in $\{1, \ldots, T\}$, i.e. no levels that could lead to values of potential outcomes that are not represented in $\{Y_i(t), i = 1, \ldots, N; t = 1, \ldots, T\}$. With the authors' notation indexing Y outcomes by treatments and mediators, the stable unit treatment value assumption implies that, given a fixed treatment level, say t, no matter how we force the mediator M to change its value for unit i from $M_i(t)$ to another value, $M^* \neq M_i(t)$, the outcome $Y_i(t, M^*)$ would remain the same, which, if implausible for any i, makes the stable unit treatment value assumption implausible and thereby makes $Y_i(t, M^*)$ functionally ill defined because of its multiple values and thus makes estimands based on the notation ill defined, as argued in Rubin (1975), page 234, Rubin (1986) and Rubin (2010), pages 40–41.

To make the stable unit treatment value assumption plausible in this case, the essential conceptual task is to formulate an assignment mechanism, not only for treatment levels, but also for mediator levels given each treatment level (Mealli and Rubin, 2003), typically either ignorable (Rubin, 1978) or latently ignorable (Frangakis and Rubin, 1999); the former relies on apposite covariates—as in Nedelman *et al.* (2003); the latter typically relies also on principal stratification (Frangakis and Rubin, 2002)—as in Jin and Rubin (2008). Ill-defined notation and the jargon of direct and indirect effects distracts us from this essential, problem-specific, conceptual task—revealed by Fisher's using such jargon to justify covariance adjustment for observed values of mediators without consideration of assignment mechanisms for them (Rubin (2005), section 7).

Marc Saez (*University of Girona, and Consortium for Biomedical Research Network in Epidemiology and Public Health, Barcelona*)

I congratulate the authors for their splendid work. I think that they contribute in an important way to investigating the explanation of causal rnechanisms. However, I am not very sure that they have succeeded, indeed, in identifying causal mechanisms. Although the theoretical argument of the two experimental designs that they propose is impeccable, the examples they provide (i.e. Sections 4.1.3 and 4.2.3) do not satisfy the same consistency assumption that is unfulfilled by the parallel and crossover designs, namely that experimental subjects need to be kept unaware of the manipulation. Of course, this does not necessarily mean that the generalization of the parallel and crossover designs by allowing for imperfect manipulation does not help to identify, effectively, average natural indirect effects but, perhaps, the choice of the examples was not successful. So I would like to ask the authors to show an example with, maybe, fewer assumptions. In any case, I think that the authors have contributed in an excellent way to establishing the theoretical foundations of the identification of causal mechanisms, particularly when it is perfectly possible to manipulate an intermediate variable.

Michael E. Sobel (*Columbia University, New York*)

I congratulate Imai, Tingley and Yamamoto for proposing creative experimental designs to help to identify pure direct and indirect effects. This is challenging because there are no observations $Y_i\{t, M_i(t')\}$ (i denotes subject, $T = t$ denotes assignment to treatment t and $M_i(t')$ is the mediator when $t \neq t'$), yet one must identify $E[Y\{t, M(t')\}]$. Identifying sequential ignorability assumptions (several are referenced in the paper) have been given, but these are typically substantively unreasonable. The authors avoid these in the parallel design by adding to the usual 'single-experiment design' a second experiment with both treatment assignment and the mediator randomized, thereby identifying controlled effects $E\{Y(t, m) - Y(t', m')\}$. Still, additional assumptions are needed to identify pure direct and indirect effects; the authors assume no interaction at the unit level. This is also very strong, and often not credible. They acknowledge this, developing sharp bounds for the parallel design that hold without this assumption. Their modified cross-

over design is nice, and the assumptions, although strong, seem more possible to meet. Similar remarks apply to the encouragement versions.

Direct and indirect effects reflect processes involving causation, providing useful information about the role of the mediator in the relationship between treatment assignment and response. But even leaving aside how one might, in the spirit of this paper, define and formalize the notion of a causal mechanism, and what it would mean to have a probabilistic causal mechanism (or should it be causal probabilistic mechanism?), it is useful to recognize that identification and estimation of direct and indirect effects need not reveal much about a causal process at work.

Consider the following hypothetical, deliberately oversimplified example. Suppose that there is a function $g(\mathbf{x}, t, m)$, where possibly $g(\mathbf{x}, t, m) = g(\mathbf{x}, t, m')$ for every $(\mathbf{x}, t)$ and (m, m'), such that $Y_i\{t, M_i(t)\} = g\{\mathbf{x}_i, t, M_i(t)\}\, M_i(t)$ and $Y_i\{t, M_i(t')\} = g\{\mathbf{x}_i, t, M_i(t')\}\, M_i(t')$ The indirect and direct effects are respectively

$$E[Y\{1, M(1)\} - Y\{1, M(0)\}] = E[g\{\mathbf{X}, 1, M(1)\}\, M(1) - g\{\mathbf{X}, 1, M(0)\}\, M(0)], \tag{19}$$

$$E[Y\{1, M(0)\} - Y\{0, M(0)\}] = E([g\{\mathbf{X}, 1, M(0)\} - g\{\mathbf{X}, 0, M(0)\}]\, M(0)). \tag{20}$$

Suppose that the authors' crossover experiment can be used to identify these effects. We can then obtain good estimates of these, with little knowledge of mechanisms: we do not know how g and M combine, nor the causal relationship between g and M, nor even that there is such a g.

The example suggests the difficulty, using even the improved experimental designs in the paper, of learning about causal mechanisms. Unless the science is already strong, it may prove very difficult to do so. That said, Imai, Tingley and Yamamoto have made a very nice contribution, and certainly a step in the right direction.

Tyler J. VanderWeele (*Harvard University, Cambridge*) **and Richard A. Emsley** (*University of Manchester*) Imai, Tingley and Yamamoto are to be congratulated for fine methodologic work which has provided experimental designs and theoretical results that together allow researchers at least sometimes to identify the sign of a mediated effect without any assumptions beyond so-called 'consistency' (see VanderWeele and Vansteelandt (2009) and VanderWeele (2012)), contrasting with prior work on bounds (Sjölander, 2009; Kaufman *et al.*, 2009; Robins and Richardson, 2010). They achieve these results by relying on fairly complex experimental designs such as when two trials are run, one in which treatment is randomized and another in which both treatment and mediator are randomized or alternatively trials in which it is possible to re-randomize, without carry-over, both treatment and mediator.

Although their designs have considerable identification power, they would, in many settings, be difficult to implement in practice. There is a trade-off between the complexity and practicality of the design on the one hand and strength of assumptions that must be employed to assess mediated effects on the other. A more common setting than the designs that they have considered is one in which treatment has been randomized in one trial, and the mediator has been randomized in another trial, possibly even with a different population from that of the first trial. The effect of treatment on the mediator and the outcome can be assessed in the first trial; the effect of the mediator on the outcome can be assessed in the second. Such designs lack the identification power of those considered by the authors and must make additional assumptions such as no interaction in expectation, cross-world independence and transportability when two different populations are used in the two experiments. But such designs would be easier to implement in practice and could even make use of existing trials and published data. We have been developing methods for such settings elsewhere (Emsley and VanderWeele, 2012). Although these methods do not allow for the identification of mediated effects without very strong assumptions, they can be useful in informing sensitivity analyses for these mediated effects. Such an approach constitutes an intermediate between the extremes of merely relying on observational studies and sensitivity analysis (Imai *et al.*, 2010; VanderWeele, 2010) or alternatively employing the complex experimental designs that were presented in the paper under discussion. However, when the parallel and crossover designs described by Imai, Tingley and Yamamoto are possible to implement, they clearly constitute a superior and more rigorous approach to assessing causal mechanisms.

The **authors** replied later, in writing, as follows.

We begin by thanking a total of more than 25 scholars from various disciplines for their valuable contributions. The fact that such a large number of contributions have been submitted reflects the interdisciplinary importance and challenges of identifying causal mechanisms. Given the limited space, we shall focus on

several common themes and reserve for future occasions our specific responses to the other points raised by each discussant.

Should scientists conduct causal mediation analysis?

Some discussants believe that the efforts to improve the credibility of causal mediation analysis may not be so worthwhile. There appear to be two main reasons for this scepticism: one fundamental and the other more practical. The fundamental criticism is that our primary estimand, the average causal mediation effect (ACME), is of limited scientific value and thus we should instead focus on some other quantity. Some contributors (e.g. Didelez and Egleston) propose as an alternative the average controlled direct effect (ACDE), defined as $\mathbb{E}\{Y_i(t,m) - Y_i(t,m')\}$. Others (e.g. Mealli and Rubin) argue for the principal strata direct effect (PSDE), such as the dissociative effect $\mathbb{E}[Y_i\{t, M_i(t)\} - Y_i\{t', M_i(t')\} | M_i(t) = M_i(t')]$.

As explained in our paper, the ACDE represents the effect of manipulating both the treatment and the mediator to specific values and thus is not directly informative of the causal process through which the treatment affects the outcome. In contrast, the ACME formalizes the notion of a causal process by considering the counterfactual outcome values which would realize when the mediators were changed as they naturally would in response to the treatment. Putting aside the terminological issue of what should be labelled a 'causal mechanism' (e.g. Sobel), scientists across disciplines very often aim to learn about causal processes. This is because scientists care not only about *changing* the world by means of external intervention, but also about *understanding* the way that the world works.

In the job market discrimination example that is discussed in our paper, social scientists are often interested in uncovering the causal process which leads an African American applicant to fewer job opportunities. Their goal is to understand the actual corporate hiring practices in the hope that such understanding will shed light on the nature of discriminatory behaviour in a society and more generally among human beings. Does discrimination arise from the perceived difference in qualifications between black and white applicants, or from the fact that the applicant is black? This is a descriptive (rather than prescriptive) causal question that can be most directly answered by quantifying the natural causal process.

In contrast, the PSDE represents the average treatment effects on the outcome among the units of specific latent characteristics defined by the potential values of the mediator. It is argued that the PSDE is preferable because the ACME is an '*a priori* counterfactual' quantity. It is argued that the PSDE avoids such pure counterfactuals yet still conveys some information about the causal mechanism of interest, because a non-zero dissociative effect implies the existence of causal pathways other than through M at least for a certain subpopulation. In our view, the conceptual difference between statements such as $Y_i\{t, M_i(t')\}$ and $M_i(t) = M_i(t')$ is less fundamental, since both are unobservable (as pointed out by Berzuini). Instead, we argue that the direct correspondence between the ACME and the concept of a causal process provides a sufficient ground for investigating this quantity. In fact, for a subpopulation with $M_i(t) = M_i(t')$, the average dissociative effect equals the average (natural) direct effect (i.e. the difference between the average treatment effect and the ACME). This close connection between the ACME and the dissociative effect makes it possible for the researcher to learn about causal processes from the PSDE (see also VanderWeele (2008)).

Of course, we do not imply that other causal quantities such as the ACDE and PSDE are of little value to scientists. In the above example, the ACDE will be more useful than the ACME if the researcher is interested in the question of whether a policy intervention to improve the qualifications of minority applicants, say through a job training programme, increases their employment prospects. We emphasize that the experimental designs that are proposed in our paper all nest the standard experiment in which only the treatment is randomized, and the parallel design in particular can point-identify the ACDE without additional untestable assumptions. Moreover, the 'augmented design' that has recently been proposed by Mattei and Mealli (2011) for the estimation of the PSDE is also nested in our parallel encouragement design. Therefore, the designs that are proposed in our paper simply expand the realm of possibility for experimental investigations into causal mechanisms. In fact, no opportunity will be lost by adopting one of our designs instead of simply conducting a standard single experiment (except the loss of statistical power, which may be an important concern in some situations as pointed out by Albert, Jo and Lange).

In the end, we believe that scientists should ultimately determine their causal quantity of interest in light of the specific applied problems they face. In our view, the job of statisticians in causal investigations are twofold:

(a) to clarify the assumptions that are required for the identification of the causal quantities that scientists wish to estimate, and
(b) to devise new methodological tools such as alternative designs and estimation techniques that help scientists to infer these quantities from the observed data better.

The choice of causal quantities should depend on the particular scientific questions being asked. It is clear that scientists are often interested in the examination of causal processes, and the ACME addresses this question most directly.

The second, more practical argument against causal mediation analysis is that, even if the ACME is of scientific interest, scientists should refrain from studying it because the identification of the ACME requires untestable assumptions, which may be difficult to justify in many applied research settings (e.g. Bullock and Green, Didelez, Draper and Glynn). In particular, concerns are raised about the plausibility of the assumptions such as consistency and the exclusion restriction. Although these assumptions should be taken seriously in applied research, we argue that the difficulty of causal mediation analysis should not be the sole reason to deter statisticians from working on related methodological problems. Typical applied research, especially in medical and social sciences, invokes several untestable assumptions. For example, the use of the instrumental variables method is usually accompanied by the assumptions of monotonicity and exclusion restriction. Even in randomized experiments, the consistency assumption (i.e. the stable unit treatment value assumption) may not be entirely valid. These concerns should not imply that empirical findings based on such assumptions are to be completely discredited. If such a perspective is applied, there will be very few valid studies left in many of the disciplines in the social and medical sciences!

A more constructive approach would be to confront these methodological challenges directly. In general, there are at least two ways in which statisticians can help scientists in this regard. First, the lack of point identification does not necessarily imply the absence of information about the ACME. As we demonstrate in the paper, the sharp bounds on the ACME can be derived to quantify precisely how much one can learn from the observed data without untestable assumptions. Indeed, some of the contributors (e.g. Ramsahai, Richardson and Robins) have taken this approach in their contributions and others have applied it in other contexts (e.g. Manski (2007)). Second, sensitivity analysis can be conducted to investigate how robust one's empirical findings are to the potential violations of such assumptions (e.g. Longford). Although we did not discuss them in our paper, several sensitivity analysis methods have already been developed for causal mediation analysis under the standard experimental design (e.g. Imai, Keele and Tingley (2010), Imai, Keele and Yamamoto (2010), VanderWeele (2010) and Tchetgen Tchetgen and Shpitser (2011)) and for some of the designs proposed in our paper (Imai and Yamamoto, 2012).

Open methodological issues and future research agenda

The methodological literature on causal mediation analysis has evolved rapidly over the last decade and we expect this trend to continue. Many of the contributors who accept the importance of causal mediation analysis suggest open methodological issues. We outline these and other challenges here in the hope that they guide future methodological research.

First, the main message of our paper is to draw attention to the 'design-based approach' to causal mediation analysis. Whereas prior research focused on various statistical methods under the standard experiment design, relatively little attention has been paid to the question of how to design randomized experiments differently to conduct causal mediation analysis with more credible assumptions. We hope that future research extends our work and develops alternative designs. Several contributors to this discussion appear to have already been moving in this direction by considering the use of covariates and other information (e.g. Albert, VanderWeele and Emsley, Hong, MacKinnon, Mateu and his colleagues and Saez; see also Section 3.1.2 of our paper). We look forward to seeing these new ideas in print. These new experimental designs are also important because they naturally serve as templates for observational studies. In Imai *et al*. (2011), we describe a couple of empirical studies in political science where the researchers analyse the observational study analogue of the crossover design that is proposed in our paper. These studies focus on the estimation of incumbency advantage, a topic which is mentioned by one of our contributors (Gelman).

Second, another important area of future research concerns multiple mediators because applied researchers are often interested in investigating the relative importance of one mediator over another (e.g. Longford). The key idea behind the proposed experimental designs is to side-step this issue by manipulating one specific mediator of interest. However, as some contributors pointed out, in practice manipulating one mechanism in isolation may be difficult, leading to the situation where multiple mediators are affected by an intervention. For this reason, it is critical to develop statistical methods that directly deal with the presence of multiple mediators. For example, Albert and Nelson (2011) discussed model-based estimation strategies for path-specific effects in the presence of multiple mediators. In Imai and Yamamoto (2012), we develop semiparametric linear models and sensitivity analyses for the potential violation of required identification assumptions concerning multiple mediators.

Third, there may be alternative approaches to causal mechanisms that are quite different from what is discussed in our paper. Some contributors mention the use of a decision theoretic framework (e.g.

Berzuini and Ramsahai). Another approach is based on the identification of sufficient causes, which is briefly discussed in our paper. These alternative approaches may shed new light on key methodological issues. For example, in his discussion, Ramsahai shows how to relax the deterministic assumptions that are made in our paper and examines the effects of doing so on the identification power of the designs proposed.

Finally, we conclude our discussion by emphasizing the importance of close collaboration between statisticians and applied researchers. As George Box succinctly put it, 'the business of the statistician is to catalyze the scientific learning process'. Any study of causal mechanisms will be best designed by taking into account specific aspects of scientific theories under investigation. Although the experimental designs that are proposed in our paper may serve as a starting point, we believe that in many situations they must be modified to address directly the methodological challenges that are faced by the researcher. In particular, practical difficulties of causal mediation analysis can be overcome by technological advances (as in the neuroscience example in our paper) and creativity on the part of the researcher (as in the labour market discrimination example). Some contributors discussed potential applications and specific challenges that range from medicine and social sciences to engineering (e.g. Egleston, Gelman, Leiva and Porcu, and Kuroki).

The challenges of causal mediation analysis should therefore motivate, rather than discourage, scientists and statisticians who are working on this important problem. For many statisticians, the mantra 'No causation without manipulation', which was put forth by Holland (1986) more than two decades ago, has been a starting point of causal analysis. Although we agree on the fundamental importance of manipulation in any causal analysis, this mantra should not be taken as a commandment that forbids certain scientific inquiry. Recently, Judea Pearl proposed another mantra 'Causation precedes manipulation'. This reminds us that manipulation is merely a tool that is used by scientists to identify causal quantities of interest. It is clear to us, and hopefully to readers, that statisticians should no longer be passively analysing the data collected by applied researchers. Rather, they must understand the causal mechanisms that are specified by scientific theories and work together with applied researchers to devise an optimal design for testing them.

References in the discussion

Albert, J. M. (2008) Mediation analysis via potential outcomes models. *Statist. Med.*, **27**, 1282–1304.

Albert, J. M. and Nelson, S. (2011) Generalized causal mediation analysis. *Biometrics*, **67**, 1028–1038.

Baron, R. M. and Kenny, D. A. (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations. *J. Personlty Socl Psychol.*, **51**, 1173–1182.

Bauer, P. and Kieser, M. (1999) Combining different phases in the development of medical treatments within a single trial. *Statist. Med.*, **18**, 1833–1848.

Cai, Z., Kuroki, M., Pearl, J. and Tian, J. (2008) Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, **64**, 695–701.

Chen, H., Geng, Z. and Jia, J. (2007) Criteria for surrogate end points. *J. R. Statist. Soc.* B, **69**, 919–932.

Clogg, C., Petkova, E. and Shihadeh, E. (1992) Statistical methods for analyzing collapsibility in regression models. *J. Educ. Statist.*, **17**, 51–74.

Cole, S. R. and Frangakis, C. E. (2009) The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, **20**, 3–5.

Cook, T. D. and Campbell, D. T. (1979) *Quasi-experimentation*. Chicago: Rand-McNally.

Dawid, A. P. (2002) Influence diagrams for causal modelling and inference. *Int. Statist. Rev.*, **70**, 161–189.

Dawid, A. P. (2003) Causal inference using influence diagrams: the problem of partial compliance. In *Highly Structured Stochastic Systems* (eds P. J. Green, N. L. Hjort and S. Richardson). New York: Oxford University Press.

Didelez, V., Dawid, A. P. and Geneletti, S. (2006) Direct and indirect effects of sequential decisions. In *Proc. 22nd Conf. Association for Uncertainty in Artificial Intelligence* (eds R. Dechter and T. Richardson), pp. 138–146. Corvallis: Association for Uncertainty in Artificial Intelligence Press.

Emsley, R. A. and VanderWeele, T. J. (2012) Mediation and sensitivity analysis using two or more trials. *Technical Report*.

Fearon, J. D. and Laitin, D. D. (2003) Ethnicity insurgency, and civil war. *Am. Polit. Sci. Rev.*, **97**, 75–90.

Frangakis, C. E. and Rubin, D. B. (1999) Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, **86**, 365–379.

Frangakis, C. E. and Rubin, D. B. (2002) Principal stratification in causal inference. *Biometrics*, **58**, 21–29.

Frumento, P., Mealli, F., Pacini, B. and Rubin, D. (2012) Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *J. Am. Statist. Ass.*, **107**, 450–466.

Geneletti, S. (2007) Identifying direct and indirect effects in a non-counterfactual framework. *J. R. Statist. Soc.* B, **69**, 199–215.
Greiner, D. J. and Rubin, D. B. (2011) Causal effects of perceived immutable characteristics. *Rev. Econ. Statist.*, **93**, 775–785.
Holland, P. W. (1986) Statistics and causal inference (with discussion). *J. Am. Statist. Ass.*, **81**, 945–960.
Hong, G. (2010) Ratio of mediator probability weighting for estimating natural direct and indirect effects. *Proc. Biometr. Sect. Am. Statist. Ass.*, 2401–2415.
Hong, G., Deutsch, J. and Hill, H. D. (2011) Parametric and non-parametric weighting methods for estimating mediation effects: an application to the National Evaluation of Welfare-to-Work Strategies. *Proc. Socl Statist. Sect. Am. Statist. Ass.*, 3215–3229.
Hoyle, R. H. and Kenny, D. A. (1999) Sample size, reliability, and tests of statistical mediation. In *Statistical Strategies for Small Sample Research* (ed. R. H. Hoyle), pp. 195–222. Thousand Oaks: Sage.
Humphreys, M. (2005) Natural resources, conflict, and conflict resolution: uncovering the mechanisms. *J. Conflct Resoln*, **49**, 508–537.
Imai, K., Keele, L. and Tingley, D. (2010) A general approach to causal mediation analysis. *Psychol. Meth.*, **15**, 309–334.
Imai, K., Keele, L., Tingley, D. and Yamamoto, T. (2011) Unpacking the black box of causality: learning about causal mechanisms from experimental and observational studies. *Am. Polit. Sci. Rev.*, **105**, 765–789.
Imai, K., Keele, L. and Yamamoto, T. (2010) Identification, inference, and sensitivity analysis for causal mediation effects. *Statist. Sci.*, **25**, 51–71.
Imai, K. and Yamamoto, T. (2012) Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. Submitted to *Polit. Anal.* (Available from `http://imai.princeton.edu/research/medsens.html`.)
Jin, H. L. and Rubin. D. B. (2008) Principal stratification for causal inference with extended partial compliance. *J. Am. Statist. Ass.*, **103**, 101–111.
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. *J. Statist. Planng Inf.*, **139**, 3473–3487.
Leiva, V., Barros, M., Paula, G. A. and Galea, M. (2007) Influence diagnostics in log-Birnbaum-Saunders regression models with censored data. *Comput. Statist. Data Anal.*, **51**, 5694–5707.
Liu, Q., Proschan, M. A. and Pledger, G. W. (2002) A unified theory of two-stage adaptive designs. *J. Am. Statist. Ass.*, **97**, 1034–1041.
MacKinnon, D. P. (2008) *Introduction to Statistical Mediation Analysis*. New York: Erlbaum.
MacKinnon, D. P. and Pirlott, A. G. (2010) The unbearable lightness of *b*: approaches to improve causal interpretation of the *M* to *Y* relation. *Society for Personality and Social Psychology Conf., Las Vegas*.
Manski, C. F. (2007) *Identification for Prediction and Decision*. Cambridge: Harvard University Press.
Mark, M. M. (1986) Validity typologies and the logic and practice of quasi-experimentation. In *Advances in Quasi-experimental Design and Analysis* (ed. A. W. M. K. Trochim), pp. 47–66. San Francisco: Jossey-Bass.
Mattei, A. and Mealli, F. (2011) Augmented designs to assess principal strata direct effects. *J. R. Statist. Soc.* B, **73**, 729–752.
Mealli, F. and Rubin, D. B. (2003) Assumptions allowing the estimation of direct causal effects. *J. Econmetr.*, **112**, 79–87.
Nedelman, J. R., Rubin, D. B. and Sheiner, L. B. (2007) Diagnostics for confounding in PK/DD models for oxcarbazepine. *Statist. Med.*, **26**, 290–308.
Neyman, J. (1990) Sur les applications de la théorie des probabilités aux experiences agricoles: essay des principes.. *Statist. Sci.*, **5**, 465–472 (Engl. tránsl.).
Pearl, J. (2001) Direct and indirect effects. In *Proc. 17th Conf. Uncertainty in Artificial Intelligence* (eds J. S. Breese and D. Koller), pp. 411–420. San Francisco: Morgan Kaufmann.
Ramsahai, R. R. (2007) Causal bounds and instruments. In *Proc. 23rd A. Conf. Uncertainty in Artifical Intelligence*, pp. 310–317. Corvallis: Association for Uncertainty in Artificial Intelligence Press.
Ramsahai, R. R. (2012) Causal bounds and observable constraints for non-deterministic models. *J. Mach. Learn. Res.*, **13**, 829–848.
Robins, J. (1986) A new approach to causal inference in mortality studies with sustained exposure periods—applications to control of the healthy worker survivor effect. *Math. Modlng*, **7**, 1393–1512.
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), pp. 70–81. Oxford: Oxford University Press.
Robins, J. M. and Richardson, T. S. (2010) Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding and Determinants of Disorders and Their Cures* (eds P. Shrout, K. M. Keyes and K. Ornstein), pp. 103–158. New York: Oxford University Press.
Robins, J. M. and Wasserman, L. (1997) Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proc. 13th Conf. Uncertainty in Artificial Intelligence* (eds D. Geiger and P. Shenoy), pp. 409–420. San Francisco: Morgan Kaufmann.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Rubin, D. B. (1975) Bayesian inference for causality: the importance of randomization. *Proc. Socl Statist. Sect. Am. Statist. Ass.*, 233–239.
Rubin, D. B. (1977) Assignment to treatment group on the basis of a covariate. *J. Educ. Statist.*, **2**, 1–26; correction, 384.
Rubin, D. B. (1978) Bayesian inference for causal effects: the role of randomization. *Ann. Statist.*, **6**, 34–58.
Rubin, D. B. (1980) Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *J. Am. Statist. Ass.*, **75**, 591–593.
Rubin, D. B. (1986) Which ifs have causal answers? *J. Am. Statist. Ass.*, **82**, 961–962.
Rubin, D. B. (2005) Causal inference using potential outcomes: design, modeling, decisions. *J. Am. Statist. Ass.*, **100**, 322–331.
Rubin, D. B. (2010) Reflections stimulated by the comments of Shadish (2009) and West & Thoemmes (2009). *Psychol. Meth.*, **15**, 38–46.
Shimizu, S. and Kano, Y. (2008) Use of non-normality in structural equation modeling: application to direction of causation. *J. Statist. Planng Inf.*, **138**, 3483–3491.
Sjölander, A. (2009) Bounds on natural direct effects in the presence of confounded intermediate variables. *Statist. Med.*, **28**, 558–571.
Smith, E. R. (1982) Beliefs, attributions, and evaluations: nonhierarchical models of mediation in social cognition. *J. Personlty Socl Psychol.*, **43**, 248–259.
Taguchi, G. (1987) *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Costs* (Engl. Transl.). New York: Quality Resources.
Tchetgen Tchetgen, E. J. and Shpitser, I. (2011) Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Working Paper 130*. Harvard University School of Public Health, Cambridge. (Available from `http://biostats.bepress.com/harvardbiostat/paper130`.)
VanderWeele, T. J. (2008) Simple relations between principal stratification and direct and indirect effects. *Statist. Probab. Lett.*, **78**, 2957–2962.
VanderWeele, T. J. (2010) Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, **21**, 540–551.
VanderWeele, T. J. (2012) Mediation analysis with multiple versions of the mediator. *Epidemiology*, **23**, 454–463.
VanderWeele, T. J. and Vansteelandt, S. (2009) Conceptual issues concerning mediation, interventions and composition. *Statist. Interface—Special Issue on Mental Health and Social Behavioral Science*, **2**, 457–468.
Wright, S. (1921) Correlation and causation. *J. Agric. Res.*, **20**, 557–585.