Ch. 5 Mediation: Exploring Relevant Mechanisms

Dylan Baker

January 29, 2025

Contents

1 Introduction

Mediation analysis seeks to answer the question "what is the causal pathway from treatment (D) to outcome (Y)?"

Running examples:

- Scurvy, Lemon Juice, and Vitamin C: Many sailors suffered from scurvy, Y. It was identified that lemon juice consumption, D, led to a decrease in scurvy. It wasn't until much later that doctors understood that the mediator was vitamin C.
- Burland, Dynarski, Michelmore, Owen, Raghuraman (2023) conduct a natural field experiment in which they compare college application rates among low-income high school students to the University of Michigan. The two treatments were to either offer free tuition to everyone or to offer free tuition to people who demonstrate need through an application process. They find that both treatments increase application rates, but the free tuition for all treatment has a larger effect (28 p.p. over control and 19 p.p. over the need-based treatment). In this case, alleviating uncertainty was a partial mediator.
- Bursztyn, González, and Yanagizawa-Drott (2020) analyzed the relationship between perceived norms and female labor force participation in Saudi Arabia. They found that most men both thought it was acceptable for women to work outside of the home and underestimated the share of other men who thought the same. Implementing the information treatment of clarifying the true share of other men who thought it was acceptable increased various outcomes related to the wives of these men working outside of the home. A partial mediator considered here is the change in perceived norms among the *wives* who did not receive the treatment, but may have heard about it from their husbands.

2 Mediation: The Basics of Causal Pathways

2.1 Complete or Partial Mediation

Mediation can be either complete or partial, in which, intuitively, the effect of D on Y either flows completely or only partially through the mediator M.

- Lemon Juice and Scurvy: The effect of lemon juice on scurvy is completely mediated by vitamin C.
- Certainty and College Applications: The effect of tuition assistance on college applications is partially mediated by alleviating uncertainty. However, a direct effect of D may also be influential. The book lists: "the effect of colorful mailings, encouragement to apply, and detailed aid information."
- Norms and Female Labor Force Participation: The effect of the information treatment on Saudi Arabian men's willingness to let their wives work via "relaxing incorrect conformity motives" is the direct effect. A partial mediator is the change in their wives' willingness to work by "relaxing incorrect conformity motives" after hearing about the treatment from their husbands.

Questions

I'm not sure that I understand at this moment why things like "colorful mailings" constitute a direct effect, rather than another mediator.

2.2 Decomposing Total Effects in the Presence of Mediators

As a simple example, suppose there is a binary treatment variable, D, and a binary mediator, M.

Define $M_i(d)$ to be the value of the mediator under treatment condition $d \in \{0, 1\}$.

Define $Y_i\left(M_i(d), d\right)$ to be the value of the outcome under treatment condition d and mediator value $M_i(d)$.

There is now no one uncontroversial definition of the ATE.

Definition D.1: Average Direct Effect (ADE)

One ATE iteration that may be of interest in this setting is the average direct effect:

$$\text{ADE}(d) \equiv \mathbb{E}\left[Y_i\left(M_i(d), 1\right) - Y_i\left(M_i(d), 0\right)\right]$$

"The ADE corresponds to the average effect of treatment once we average over the values of the mediator that arise naturally in the population."

Definition D.2: Average Indirect Effect (AIE)

In a similar spirit, we can define the average indirect effect:

$$\text{AIE}(d) \equiv \mathbb{E}\left[Y_i\left(M_i(1), d\right) - Y_i\left(M_i(0), d\right)\right]$$

which corresponds to the average effect when we hold the treatment condition fixed.

In the case of complete mediation, the ADE is zero, and the AIE holds the total effect.

Example 2.1. In the Saudi Arabian norms experiment, "AIE(0) is capturing how much higher (or lower) female labor force participation would be had no husbands received the norms information, but had one group had their wives conveyed information as if their husbands had received norms information, isolating the potential indirect effect."

End of Example.

What does our typical ATE capture in this binary mediator scenario?

Notice that you can re-write the ATE as follows:

$$
\begin{aligned}
\text{ATE} =& \mathbb{E}\left[Y_i\left(M_i(1),1\right) - Y_i\left(M_i(0),0\right)\right] && (1)\\
=& \underbrace{\mathbb{E}\left[Y_i\left(M_i(1),1\right) - Y_i\left(M_i(1),0\right)\right]}_{\text{ADE}(1)} + \underbrace{\mathbb{E}\left[Y_i\left(M_i(1),0\right) - Y_i\left(M_i(0),0\right)\right]}_{\text{AIE}(0)} && (2)\\
=& \underbrace{\mathbb{E}\left[Y_i\left(M_i(1),1\right) - Y_i\left(M_i(0),1\right)\right]}_{\text{AIE}(1)} + \underbrace{\mathbb{E}\left[Y_i\left(M_i(0),1\right) - Y_i\left(M_i(0),0\right)\right]}_{\text{ADE}(0)}
\end{aligned}
$$

Then, the ATE is a sum of the ADE and AIE under different treatment conditions.

2.3 Moving the Goalposts: Controlled and Principal-Strata Effects

Suppose that we can control both the treatment and mediator conditions. Then, we can manipulate each in what functionally amounts a "full-factorial design in the space of $D \times M$."

Definition D.3: Average Controlled Direct Effect (ACDE)

What we may have previously called an interaction effect, we now "re-interpret" to as the average controlled direct effect:

$$\text{ACDE}(m) \equiv \mathbb{E}\left[Y_i(m,1) - Y_i(m,0)\right]$$

In practice, this may be hard to attain. For one thing, it may be the case that varying the mediator is simply not possible for ethical, legal, or practical reasons. Moreover, if treatments and mediators endogenously interact, then the level of the mediator imposed by the researcher may differ from the level that would have arisen naturally. In that case, it may be that $\text{ACDE}(1) \neq \text{ADE}(1)$. That is, the "controlled" effect may differ from the "organic" effect, because the mediator may take on a different level when controlled compared to when it organically emerged as a result of the treatment. This places a responsibility on the researcher to think carefully and choose practically interesting levels of the mediator.

2.3.1 "Always"-Mediator-Takers

For the subset of the population that always takes the mediator, i.e., $M(1) = M(0) = 1$, the sub-population ATE is the same as the sub-population ADE, so we can get:

$$
\begin{aligned}
\text{subpopulation ATE} &= \mathbb{E}\left[Y_i\left(M_i(1),1\right) - Y_i\left(M_i(0),0\right) \mid M_i(1) = M_i(0) = 1\right] && \text{From (1)}\\
&= \underbrace{\mathbb{E}\left[Y_i(1,1) - Y_i(1,0)\right]}_{\text{ADE}(1)} + \underbrace{\mathbb{E}\left[Y_i(1,0) - Y_i(1,0)\right]}_{\text{AIE}(0)=0} && \text{From (2)}
\end{aligned}
$$

3 Applied Mediation Analysis for Economic Experts

3.1 A Parametric Workhorse and its Pitfalls

"Up to this point, we have focused on discussing general mediation parameters of interest, without introducing functional form assumptions."

"Consider the following system of linear equations with constant coefficients:"

$$Y_i = \mu + \lambda_{dy} D_i + \lambda_{my} M_i + X_i' \delta + \epsilon_i$$
$$M_i = \alpha + \lambda_{dm} D_i + X_i' \gamma + v_i$$

See Figure 1 for a graphical representation of this system.

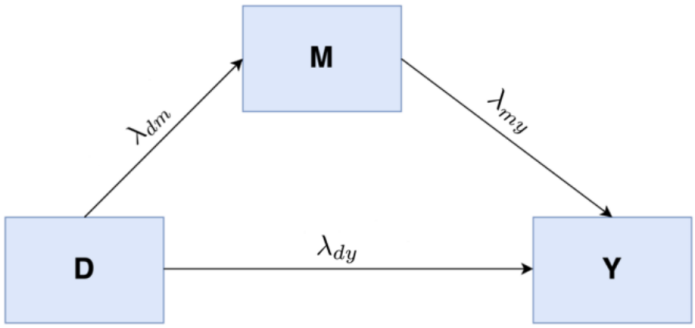

Figure 1: Mediation Graph with Linear Functional Form

We can then see from either the model or the graph that the average indirect effect is given by $\lambda_{dm}\lambda_{my}$. That is, we're scaling the effect of M on Y by how much D affects M.

This result has inspired many papers to engage in 2-stage experiments in which they first randomize D and estimate the effects on M and then, in a second experiment, randomize D and measure its effect on Y controlling for M.

See the appendix of John's book for a demonstration of this.

3.2 Basic Case: Binary Randomized Treatment

There are alternative approaches available when the experimenter is able to directly manipulate M.

Given that D has been randomly assigned, we can be confident in the assumption:

Assumption: Statistical Independence of the Treatment

$$\{Y_i(1,1), Y_i(1,0), Y_i(0,1), Y_i(0,0), M_i(1), M_i(0)\} \perp D_i$$

We can then consider another assumption:

Assumption: Conditional Independence of the Mediator

$$\{Y_i(1,1), Y_i(1,0), Y_i(0,1), Y_i(0,0)\} \perp M_i \mid D_i$$

"Concretely, this assumption requires that the value of the mediator is as good as randomly assigned, even though the researcher did not have direct control over its level."

Putting these together:

Assumptions Combined: Sequential Randomization or Sequential Ignorability

Researchers typically consider the above two assumptions jointly as the assumption of sequential randomization or sequential ignorability:

$$\{Y_i(1,1), Y_i(1,0), Y_i(0,1), Y_i(0,0)\} \perp M_i \mid D_i$$
$$\{Y_i(1,1), Y_i(1,0), Y_i(0,1), Y_i(0,0), M_i(1), M_i(0)\} \perp D_i$$

Assumption: Support

An additional assumption is

$$1 > \mathbb{P}\left[D_i = 1 \mid M_i = m\right] > 0 \text{ for all } m$$

That is, for all values of the mediator, there is a positive probability of receiving the treatment.

Under these 3 assumptions, we can identify all of the 4 parameters of interest.

For example, under these assumptions, $\mathbb{E}\left[Y_i\left(M_i(1), 0\right)\right]$, the average outcome if the mediator were the value under treatment but treatment was set to 0, is given by:

$$\mathbb{E}\left[Y_i\left(M_i(1), 0\right)\right] = \mathbb{E}\left[Y_i \cdot (1 - D_i) \cdot \frac{1}{\mathbb{P}\left[D_i = 1\right]} \left(\frac{1}{1 - \mathbb{P}\left[D_i = 1 \mid M_i = m\right]} - 1\right)\right]$$

3.2.1 The Assumption that Fails

However, in practice, the assumption that is likely to fail is the assumption of conditional independence of the mediator. Realistically, the mediator value probably reflects choice and optimization by the individual, so it's unlikely that the mediator is as good as randomly assigned given treatment. E.g., in the Saudi Arabian norms experiment, this would fail in a world where whether husbands communicate with their wives about the norms is at least partially informed by how likely the information would be to influence their wives' behavior.

Questions

Verify that what I wrote in this example is correct.

3.3 Separate Randomization of Treatment and Mediator

Another common approach is to conduct 2 experiments where in the first, the experimenter randomizes D and estimates the effect on M, and in the second, the experimenter randomizes M and estimates the effect on Y.

While this method offers some intuitive appeal, it fails to recover parameters of interest, such as the AIE. See Figure 2 below, which is included in John's book and is a reproduction of a table from Imai et al. (2011).

In this example, we see that one gets a positive effect of D on M and a positive effect of M on Y: 0.2 for each. However, the causal mediation effect is actually negative: -0.2. Why is this? The issue lies in which members of the population are affected in each case. The positive effect of D on M is driven by the sub-population in the first row, those with $M_i(0) = 0$ and $M_i(1) = 1$. However, this is the exact population for whom the mediator has a negative effect on Y, i.e., for whom $Y_i(t, 0) = 1$ and $Y_i(t, 1) = 0$. Thus, the issue comes from not appreciating that the impact of D on M is not applied uniformly across the population, and it may be applied to a sub-population for whom the effect of M on Y doesn't match the average effect across the population.

Population Proportion	Potential Mediators and Outcomes				Treatment Effect on Mediator	Mediator Effect on Outcome	Causal Mediation Effect
	$M_i(1)$	$M_i(0)$	$Y_i(t, 1)$	$Y_i(t, 0)$	$M_i(1) - M_i(0)$	$Y_i(t, 1) - Y_i(t, 0)$	$Y_i(t, M_i(1)) - Y_i(t, M_i(0))$
0.3	1	0	0	1	1	−1	−1
0.3	0	0	1	0	0	1	0
0.1	0	1	0	1	−1	−1	1
0.3	1	1	1	0	0	1	0
Average	0.6	0.4	0.6	0.4	0.2	0.2	−0.2

This table, which is reproduced from Table 1 in Imai et al. (2011), shows that combining the ATEs obtained from separate randomization of the treatment (D) and mediator (M) is not sufficient to recover the average indirect effect (AIE). By first randomizing D to recover its effect on M, and then M to recover its effect on Y, we find both effects to be on average positive at 0.2. However, the AIE is negative 0.2 because the units for which the effect of D on M is positive are those for whom the effect of M on Y is negative.

Figure 2: Separate Randomization of Treatment and Mediator Failing to Recover AIE

Such an issue would be ruled out by the Sequential Ignorability assumption had it applied here. In particular, in addition to treatment D_i randomization, the assumption would mean that conditional on D_i, M_i is independent of the potential outcomes of Y_i, but this clearly doesn't hold in this example, since conditional on D_i, M_i is related to the potential outcomes of Y_i.

This means that randomizing M_i in the second experiment doesn't reflect what actually happens to M_i when we introduce D_i.

To-Do

Add a clearer explanation of showing how they're related.

3.4 Paired Design

An alternative to separate randomization is the paired design. Under the paired design, the researcher runs 2 experiments:

- Experiment A: Randomize over D.
  - Recovers the ATE, which is composed of direct and indirect effects.
- Experiment B: Randomize over D and M.
  - Recovers the ACDEs.

Experiment B can give some indication of whether there is a meaningful interaction effect: $D \times M$.

If there is not, this may be some suggestive evidence that D and M work in predominantly additively separable ways. This implies that there is only one direct and one indirect effect.

That is,

$$\text{ADE}(0) = \text{ADE}(1) \equiv \text{ADE} \tag{3}$$
$$\text{AIE}(0) = \text{AIE}(1) \equiv \text{AIE} \tag{4}$$

Moreover,

$$\text{ACDE}(0) = \text{ACDE}(1) \equiv \text{ACDE}$$

and

$$\text{ACDE} = \text{ADE} \tag{5}$$

since the idea is that ACDE isn't varying with M and hence recovers the ADE.

Moreover, the average indirect effect can be recovered as the difference-in-differences across experiments. That is:

$$\text{AIE} = \underbrace{\mathbb{E}\left[Y_i\left(M_i(1), 1\right) - Y_i\left(M_i(0), 0\right)\right]}_{\text{ATE (Experiment A)}} - \underbrace{\mathbb{E}\left[Y_i(1,1) - Y_i(0,1)\right]}_{\text{ACDE (Experiment B)}}$$

since:

$$\begin{aligned} &\text{ATE} = \text{ADE} + \text{AIE} && \text{by (2), (3), and (4)} \\ \Rightarrow &\text{AIE} = \text{ATE} - \text{ADE} && \\ \Rightarrow &\text{AIE} = \text{ATE} - \text{ACDE} && \text{by (5)} \end{aligned}$$

Thus, in this case, "all mechanism-specific components of the ATE are recovered."

3.5 Cross-Over Design

3.5.1 Setup

In the cross-over design, the researcher considers a single experiment over multiple time periods.

In the initial round, denoted with the subscript 1, the researcher randomizes D_{i1} and measures the (endogenous) mediator, M_{i1}, as well as the outcome, Y_{i1}. In the subsequent round, denoted by subscript 2, the researcher flips the treatment, so that

$$D_{i2} = 1 - D_{i1} = \begin{cases} 1 & \text{if } D_{i1} = 0 \\ 0 & \text{if } D_{i1} = 1 \end{cases}$$

but holds the mediator value fixed at the value it took in the first round, i.e., $M_{i2} = M_{i1}$. Then, the researcher measures the outcome, Y_{i2}.

3.5.2 Assumptions

- Causal Transience: Our usual causal transience assumption can be stated as:

  $$Y_{it}\left(D_{it}, \boldsymbol{D}_i, T\right) = Y_{it}\left(D_{it}, T\right)$$

  where

  $$\boldsymbol{D}_i = \left(D_{i1}, \ldots, D_{iT}\right)$$

  and in this case, $T = 2$.

- Effect of Mediator Transience: We must now also assume that fixing the value of the mediator to the initial (endogenous) value in subsequent rounds yields the same outcome as it did when it was endogenously chosen in the first round. That is, when $M_i(d) = m$:

  $$\mathbb{E}\left[Y_{i1}\left(M_i(d), d\right)\right] = \mathbb{E}\left[Y_{i2}(m, d)\right] \text{ for all } d, m.$$

- Other stuff from the within-participants design chapter presumably applies here as well.

3.5.3 Results

If our assumptions are met, we get the nice reward of being able to separately identify the AIE(d) and ADE(d) values.

As one example of how we can do this, consider:

$$
\begin{aligned}
\text{AIE}(1) \equiv & \mathbb{E}\left[Y_{i2}\left(M_i(1),1\right) - Y_{i2}\left(M_i(0),1\right)\right] \\
= & \mathbb{E}\left[Y_{i2}\left(M_i(1),1\right)\right] - \mathbb{E}\left[Y_{i2}\left(M_i(0),1\right)\right] \\
= & \mathbb{E}\left[Y_{i1}\left(M_i(1),1\right)\right] - \mathbb{E}\left[Y_{i2}\left(M_i(0),1\right)\right] \qquad \text{by assumption} \\
= & \underbrace{\mathbb{E}\left[Y_{i1} \mid D_{i1}=1\right]}_{\text{Initial period mean of initially-treated}} - \underbrace{\mathbb{E}\left[Y_{i2} \mid D_{i1}=0\right]}_{\text{Subsequent period mean of initially-untreated}}
\end{aligned}
$$

Basically, we're leveraging:

- that in the initial period, we got the average value of the outcome under treatment, including permitting the mediator to endogenously adjust, giving us an estimate of $\mathbb{E}\left[Y_{i1} \mid D_{i1}=1\right]$
- and that in the second period, we're able to hold fixed the endogenously selected value of the mediator among those untreated in period 1 and introduced treatment in period 2[1]

which, under our assumptions, gives us the AIE(1) by the above set of equalities.

Similarly,

$$
\begin{aligned}
\text{AIE}(0) \equiv & \mathbb{E}\left[Y_{i2}\left(M_i(1),0\right) - Y_{i2}\left(M_i(0),0\right)\right] \\
= & \mathbb{E}\left[Y_{i2}\left(M_i(1),0\right)\right] - \mathbb{E}\left[Y_{i2}\left(M_i(0),0\right)\right] \\
= & \mathbb{E}\left[Y_{i2}\left(M_i(1),0\right)\right] - \mathbb{E}\left[Y_{i1}\left(M_i(0),0\right)\right] \qquad \text{by assumption} \\
= & \underbrace{\mathbb{E}\left[Y_{i2} \mid D_{i1}=1\right]}_{\text{Subsequent period mean of initially-treated}} - \underbrace{\mathbb{E}\left[Y_{i1} \mid D_{i1}=0\right]}_{\text{Initial period mean of initially-untreated}}
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\text{ADE}(1) \equiv & \mathbb{E}\left[Y_{i2}\left(M_i(1),1\right) - Y_{i2}\left(M_i(1),0\right)\right] \\
= & \mathbb{E}\left[Y_{i2}\left(M_i(1),1\right)\right] - \mathbb{E}\left[Y_{i2}\left(M_i(1),0\right)\right] \\
= & \mathbb{E}\left[Y_{i1}\left(M_i(1),1\right)\right] - \mathbb{E}\left[Y_{i2}\left(M_i(1),0\right)\right] \qquad \text{by assumption} \\
= & \underbrace{\mathbb{E}\left[Y_{i1} \mid D_{i1}=1\right]}_{\text{Initial period mean of initially-treated}} - \underbrace{\mathbb{E}\left[Y_{i2} \mid D_{i1}=1\right]}_{\text{Subsequent period mean of initially-treated}}
\end{aligned}
$$

and

[1] To be super explicit, I'm saying that we're not allowing endogenous adjustment of the mediator for this second-period treated group.

$$
\begin{aligned}
\text{ADE}(0) \equiv & \mathbb{E}\left[Y_{i2}\left(M_i(0),1\right) - Y_{i2}\left(M_i(0),0\right)\right] \\
= & \mathbb{E}\left[Y_{i2}\left(M_i(0),1\right)\right] - \mathbb{E}\left[Y_{i2}\left(M_i(0),0\right)\right] \\
= & \mathbb{E}\left[Y_{i2}\left(M_i(0),1\right)\right] - \mathbb{E}\left[Y_{i1}\left(M_i(0),0\right)\right] && \text{by assumption} \\
= & \underbrace{\mathbb{E}\left[Y_{i2} \mid D_{i1}=0\right]}_{\text{Subsequent period mean of initially-untreated}} - \underbrace{\mathbb{E}\left[Y_{i1} \mid D_{i1}=0\right]}_{\text{Initial period mean of initially-untreated}}
\end{aligned}
$$

4 Appendix