Nguồn: http://egap.org/methods-guides/10-things-mechanisms

10 Things to Know About Mechanisms

- 1 Mechanisms are pathways through which X causes outcome Y.
- 2 While we don't *need* to know the mechanism to conclude that X causes Y, there are several reasons why we *want* to.
- 3 But it is extremely challenging to identify causal mechanisms because the mechanisms themselves are not randomly assigned…
- 4 …and because treatment effects are rarely homogeneous.
- 5 Many studies try to decompose a total treatment effect into its "direct" and "indirect" effects.
- 6 But be cautious about using regression analysis to decompose effects.
- 7 Sometimes subgroup analysis can provide suggestive evidence for or against a mechanism.
- 8 We can also look for suggestive evidence by looking at the effects of our treatment on various outcomes.
- 9 Designing complex treatments can help narrow our understanding of what part of the treatment is "doing the work."
- 10 Despite the difficulties in empirically measuring mechanisms, it is worth paying serious attention to them but being cautious in our language.

As social scientists, we are fascinated by causal questions. As soon as we learn that X causes Y, we want to better understand *why* X causes Y. This guide explores the role of "mechanisms" in causal analysis and will help you to understand what kinds of conclusions you may draw about them.

1 Mechanisms are pathways through which X causes outcome Y.

Mechanisms have long been at the heart of medicine. Every time a doctor prescribes a treatment, she does so out of an understanding of which chemical or physical factors cause a disease, and she prescribes a treatment that is effective because it interrupts these factors. For example, many clinical psychologists recommend exercise to patients dealing with depression. Exercise raises endorphins in the body's chemistry, which trigger positive feelings and also act as analgesics, which reduce the perception of pain. Endorphins, therefore, are a mechanism by which exercise helps reduce depression. Exercise may have positive effects on a number of other dependent variables (e.g. heart disease) through other mechanisms (e.g. elevating heart rates), but the mechanism that causes it to affect depression in particular is endorphins. We could also conclude that another treatment, such as a drug that raised endorphins, may have similar effects on depression.

Mechanisms are just as important for social sciences. Take, for example, recent research that has connected climate change to an increase in civil conflict. One study[1] claims to identify the causal effect of climate shocks on violent conflict by studying the rate of civil conflict in El Nino-affected countries during El Nino versus non-El Nino years. Suppose this study is correct. Why would experiencing a climate shock cause a country to have elevated levels of conflict? One mechanism could be poverty: climate shocks hurt the economy, and

with lower opportunity costs, individuals are more inclined to join armed groups. An alternative mechanism is physiological: people are physically wired to be more aggressive in hotter temperatures. Perhaps the mechanism is migration: climate shocks displace people in coastal regions, and this produces social conflict between migrants and natives. In reality, several or all of these mechanisms (as well as others not listed here) could be operating simultaneously, even in the same case! In many of the most interesting social science questions, there are several channels ("M"s) that could transmit the total effect of X on Y.

2 While we don't *need* to know the mechanism to conclude that X causes Y, there are several reasons why we *want* to.

In the climate/conflict example above, we can have full confidence in the researchers' ability to causally identify that climate shocks cause conflict, and yet have no evidence of which mechanism(s) is/are at work. But social scientists are interested in learning about mechanisms because they tightly relate to social science theories. For example, the "poverty" mechanism above closely relates to Gurr's[2] theory that individuals rebel when their opportunity costs of conflict are low, whereas the "migration" mechanism could support a theory of conflict based on grievances between social groups. It is no wonder that upon learning that X causes Y, social scientists immediately ask what the mechanism is – they want to relate this finding to theory!

Understanding mechanisms has not only theoretical but practical benefits. First, knowing M allows us to guess for which populations X will lead to Y. If the mechanism for climate/conflict is physiological response to heat, then climate shocks may produce conflict only when temperature is quite warm. Second, knowing M helps us to consider other outcomes that may be affected by X. If the mechanism for climate/conflict is migration, then we might also expect climate shocks to result in overuse of public goods in urban areas. Third, knowing M helps us to consider other ways to cause or avoid causing changes in Y. If the mechanism for climate/conflict is poverty, then development programs could decrease conflict by reducing the sensitivity of incomes to climate shocks, even though they can't change climate shocks.

3 But it is extremely challenging to identify causal mechanisms because the mechanisms themselves are not randomly assigned…

Consider an experimental example. Chong et al. (2015)[3] used a field experiment to study the effect of corruption information on voter turnout. They randomly assigned some polling precincts in Mexico to receive information about the corrupt use of funds within that municipality. Surprisingly, they found that treated precincts turned out to vote at lower rates than control precincts. They suggest the following mechanism at work: corruption

information convinces voters that the municipality is so severely corrupt that electing a good politician will not change it, so individuals find their vote to carry less value.

In short, their argument is:[4]

Receiving corruption information (X) −−→+→+ Believes corruption too severe (M) −−→+→+ Stay home (Y)

Chong et al. face a common obstacle in interpreting their results: their proposed mechanism was not randomly assigned. Some people are more inclined to believe that "all politicians are rascals" while others have a tendency to push for "change we can believe in." Unfortunately, we can observe only the random treatment an individual received and their non-random belief about corruption; we can't tell what belief about corruption they *would have had* if they had received the other treatment condition. This makes it impossible for us to determine, for each individual, the extent to which her decision to turn out to vote was caused by the proposed mechanism versus other mechanisms.

Some researchers try to get around this problem by estimating the *average* effect of the treatment on the mechanism and then estimating the *average* effect of the mechanism on the outcome. One reason that this is problematic is that we can imagine several factors other than the treatment that could be causing both M and Y. Suppose that the level of apathy – let's call it Q – varies among the citizens in our study, and Q has a very strong effect on both M and Y. Highly apathetic individuals might be more likely to believe that problems are beyond solving, and they might also be more likely to stay at home on election day. We are therefore likely to observe a strong correlation between M and Y that is driven by the confounder of Q, not by our treatment X. Mechanically, our results will be biased in favor of finding evidence of X's effect on Y via M simply because Q has produced a relationship between M and Y.

4 …and because treatment effects are rarely homogeneous.

The other problem with trying to decompose the average effects of X on M and then M on Y is that this approach assumes that every subject responds to the treatment identically. Recalling our example in which X is the information treatment, M is the belief that corruption is too severe, and Y is staying home, we can imagine two types of respondents. Type A thought that corruption was too severe to ever solve until she received a postcard containing information about corruption in her district. She was surprised to see that the problem was not as bad as she had expected. Formally, for Type A, M(X=0)=1M(X=0)=1 and M(X=1)=0M(X=1)=0, so X has a *negative* effect on M. Type B thought corruption was a manageable problem until she received a postcard containing information about corruption in her district. She was surprised by how extensive the problem was and gave up hope of solving the problem. Formally, for Type B, M(X=0)=0M(X=0)=0 and M(X=1)=1M(X=1)=1, so X has a *positive* effect on M. If we were to average the effects for these two types, we would see no relationship between X and M.

Type	X (Info treatment)	M conditional on X=0 (unobserved)	M conditional on X=1 (observed)	Effect of X on M	Effect of M on Y	Y (Stay home)
A	1	1	0	negative	negative	1
B	1	0	1	positive	positive	1

Estimating the role of M can be further complicated when the relationship between M and Y is also heterogeneous. Imagine that Type A only votes when she's angry (in other words, M has a *negative* effect on Y). Type A was planning on voting to express her anger over the pervasiveness of corruption in her district, even though she knew it would not have changed anything, until she learned that corruption was not as bad as she had expected it to be. Her fiery passion gone, she chooses to stay home on election day. However, Type B only votes when she thinks her vote can make a difference (in other words, M has a *positive* effect on Y). Type B was going to vote for the non-corrupt politicians in her district until she learned that they were all corrupt. Without any hope of changing the situation, she also decided to stay home on election day. For both Type A and Type B, there is an "indirect effect" of M (in other words, X affects Y through M). But we will miss this relationship in the aggregate because we will be unable to obtain unbiased estimates of the average effect of X on M.[5]

We can imagine many more "types" than just A and B – the point here is to demonstrate intuitively that because M is not randomly assigned, and because it is unlikely that the effects of X on M *and* M on Y are identical for everyone, it will be very difficult to accurately characterize how much of our effect is mediated through M.

5 Many studies try to decompose a total treatment effect into its "direct" and "indirect" effects.

Because learning about mechanisms holds such rich theoretical promise, researchers would love to quantify how much of an effect of X on Y operates via M. Sometimes researchers will try to do this through a technique called "decomposition of effects."

A decomposition of effects analysis tries to decompose a *total* effect of X on Y into the effect X has on Y *directly* and the effect of X on Y that occurs *indirectly* through M. The "total effect" refers to the Average Treatment Effect (ATE), which is simply the average effect that X has on Y. Any experiment that randomly assigns a treatment in order to observe its effects on some outcome is estimating the ATE. Next, the researcher tries to quantify the size of the effect that X has on Y through the mechanism M. This is often known as the "indirect effect" – because X is affecting Y indirectly through M – or the Average Causally Mediated Effect (ACME). Finally, the researcher will try to estimate the effect of X on Y that doesn't go through M. This is known as the "direct effect" of X on Y or the Average Controlled Direct Effect (ACDE), because it is the effect of X on Y when we control for the work that M is doing.

6 But be cautious about using regression analysis to decompose effects.

Although commonly used, using mediation regression analysis presupposes some strong and often unrealistic assumptions. We will use some code to illustrate what this method entails and demonstrate the conditions under which it can produce biased estimates.

The basic idea is that if we have data on the treatment an individual received (X), whether they exhibit the proposed mechanism (M), and what the outcome is (Y), then we can distinguish these effects using the following three regressions.

1. $M_i=\alpha_1+aX_i+e_{1i}Mi=\alpha1+aXi+e1i$
2. $Y_i=\alpha_2+cX_i+e_{2i}Yi=\alpha2+cXi+e2i$
3. $Y_i=\alpha_3+dX_i+bM_i+e_{3i}Yi=\alpha3+dXi+bMi+e3i$

How would we do this? Using equation 1, we can regress M on X to obtain the direct effect of X on M, which is the coefficient aa. Next, we turn to equation 3, in which we regress Y on M and X. In this regression, the coefficient bb represents the direct effect of M on Y when we control for X. A decomposition of effects analysis would multiply $a*ba*b$ to reveal the indirect effect of X on Y via M. To find the direct effect of X on Y, we need look no further than dd, which is the coefficient on X in equation 3 when we control for M. In other words, dd is the effect of X on Y that does not go through M. If we add the indirect effect and direct effect, we will come up with the "total effect" of X on Y, which is equal to cc. To summarize, the decomposition of effects analysis ostensibly disaggregates the total effect into the effect that is mediated via M and the effect that is not mediated via M, enabling the researcher to conclude how important M is for explaining the relationship between X and Y.[6]

The problem is that this arithmetic only works under some very strong assumptions. One of these assumptions is that the error terms in regressions 1 and 3 are unrelated to each other—in other words, M can't be predicted by unobservable factors that also predict Y. We described this problem intuitively in point 3 when we introduced Q, a confounding variable that contributes both to M and to Y and therefore engenders a very strong relationship between them, even if X's effect on Y is not operating through M at all. Now let's describe this problem using a simulation.

In the following code, we start by creating this Q variable for each individual and defining the "true" effects of X on M, M on Y, and X on Y. Next, we create hypothetical potential outcomes for M—that is, for each individual, we define what value of M they would reveal if they were treated, and what value of M they would reveal if they were untreated. These values are related not only to the "true" effect of X on M, but also to Q. Then we can also define hypothetical potential outcomes for Y. We do this for four scenarios, all of which assume constant effects, an assumption we will relax later. Two of these are simple potential outcomes of Y: the Y exhibited by the individual who is untreated and reveals her untreated M potential outcome, and the Y exhibited by the individual who is treated and reveals her treated M potential outcome. However, we also define two complex potential outcomes of Y: the Y exhibited by the individual who is untreated but reveals her treated M potential outcome, and the Y exhibited by the individual who is treated but reveals her untreated M potential outcomes. While these potential outcomes bend the mind a bit, they are important

to define in the hypothetical so that we can calculate the “true” (but inherently unobservable) direct and indirect effects to compare our decomposition analysis to.

In the second half of the code, we conduct a random assignment of treatment and proceed with the decomposition of effects analysis described above, using the data we “observe.” Under the (strong) assumptions that the error terms are uncorrelated and effects are constant across subjects, a∗ba∗b = ACME, dd = ACDE, and cc = ATE. However, the simulation reveals that a∗ba∗b > ACME and dd < ACDE; that is, we overestimated the average indirect or mediated effect (ACME) and underestimated the average direct effect (ACDE). Our decomposition of effects analysis was biased because the first assumption – uncorrelated error terms – did not hold: unobserved variable Q predicted both M and Y, and this led us to overestimate the role of the mechanism M.
Hide

```
rm(list = ls())


set.seed(20160301)


N <- 1000000


# Simulate Data, Create Potential Outcomes, Estimate "True" Effects ---------
--------------


# build in an ideosyncratic unobserved characteristic

Q_i <- rnorm(N)


# create the "true model" by defining our treatment effects (tau)

tau_X_on_M <- 0.2 # X's effect on M

tau_M_on_Y <- 0.1 # M's effect on Y

tau_X_on_Y <- 0.5 # total effect of X on Y (ATE), both through M and not thro
ugh M


# build the potential outcomes (POs) for the mediator

# individual reveals M_1 if treated; M_0 if untreated

# M is a function of both treatment and the unobserved characteristic

M_0 <- 0 * tau_X_on_M + Q_i

M_1 <- 1 * tau_X_on_M + Q_i


# we can estimate the unbiased Average Treatment Effect (ATE) of X on M

ATE_M <- mean(M_1 - M_0)

ATE_M
```

```
[1] 0.2
```

Hide

```
# build POs for the outcome variable
Y_M0_X0 <- tau_M_on_Y * (M_0) + tau_X_on_Y * 0 + Q_i
Y_M1_X1 <- tau_M_on_Y * (M_1) + tau_X_on_Y * 1 + Q_i
Y_M0_X1 <- tau_M_on_Y * (M_0) + tau_X_on_Y * 1 + Q_i # this is a "complex" PO
Y_M1_X0 <- tau_M_on_Y * (M_1) + tau_X_on_Y * 0 + Q_i # this is a "complex" PO
# some of these POs are "complex" because we are imagining what Y we would
# observe if we assigned treatment but observed the untreated M PO or
# if we assigned control but observed the treated M PO
# building these complex POs is necessary for estimating the "true" direct an
d indirect effects


# we can estimate the unbiased Average Causally Mediated Effect (ACME)
# we estimate the effects of M holding X constant
# they are the same
# this is the "indirect effect"
ACME_X0 <- mean(Y_M1_X0 - Y_M0_X0)
ACME_X1 <- mean(Y_M1_X1 - Y_M0_X1)
ACME <- mean(((Y_M1_X1 - Y_M0_X1) + (Y_M1_X0 - Y_M0_X0)) / 2)


# we can estimate the unbiased Average Controlled Direct Effect (ACDE)
# we estimate the effects of X holding M constant
# they are the same
# this is the "direct effect"
ACDE_M0 <- mean(Y_M0_X1 - Y_M0_X0)
ACDE_M1 <- mean(Y_M1_X1 - Y_M1_X0)
ACDE <- mean(((Y_M0_X1 - Y_M0_X0) + (Y_M1_X1 - Y_M1_X0)) / 2)


# now we build the simple POs for Y
Y_1 <- tau_M_on_Y * (M_1) + tau_X_on_Y * 1 + Q_i
Y_0 <- tau_M_on_Y * (M_0) + tau_X_on_Y * 0 + Q_i


# we estimate the true ATE of X on Y
# this is the "total effect"
ATE <- mean(Y_1 - Y_0)
```

```
ATE
```

```
[1] 0.52
```

Hide

```
ACDE + ACME # note that the direct and indirect effects sum to the total
```

```
[1] 0.52
```

Hide

```
ACDE
```

```
[1] 0.5
```

Hide

```
ACME
```

```
[1] 0.02
```

Hide

```
ATE_M
```

```
[1] 0.2
```

Hide

```
# Random Assignment, Revelation of POs, Attempt to Decompose Effects --------
-------------------------


# we assign half of our sample to treatment and half to control
X <- sample(c(rep(1, (N / 2)), rep(0, (N / 2))))
# we reveal POs for M and Y based on treatment assignment
M <- X * M_1 + (1 - X) * M_0
Y <- X * Y_1 + (1 - X) * Y_0


model1 <- lm(M ~ X)
a <- coef(model1)[2] # extract the coefficient to get the effect of X on M
a
```

```
        X
0.2001291
```

Hide

```
model2 <- lm(Y ~ X)
```

```
c <- coef(model2)[2] # extract the coefficient to get the total effect of X o
n Y

c
```

```
       X
0.520142
```

Hide

```
model3 <- lm(Y ~ X + M)

d <- coef(model3)[2] # extract this coefficient to get the effect of X on Y c
ontrolling for M

b <- coef(model3)[3] # extract this coefficient to get the effect of M on Y c
ontrolling for X


# some would now multiply the average effect of X on M and the average effect
of M on Y to get the average indirect/mediated effect of X on Y via M (ACME)

a * b
```

```
       X
0.220142
```

Hide

```
# but when we compare this to the true ACME, we see that this is biased

ACME
```

```
[1] 0.02
```

Hide

```
# some would also interpret the average effect of X on Y controlling for M as
the average controlled direct effect (ACDE)

d
```

```
  X
0.3
```

Hide

```
# but when we compare this to the true ACDE, we see that this is biased

ACDE
```

```
[1] 0.5
```

Hide

```
# note that we have OVERestimated the average indirect effect and UNDERestima
ted the average direct effect
```

```
# the estimates that are unbiased are the average effects of X on Y and X on
M because X is randomly assigned

a
```

```
         X
0.2001291
```

Hide

```
ATE_M
```

```
[1] 0.2
```

Hide

```
c
```

```
        X
0.520142
```

Hide

```
ATE
```

```
[1] 0.52
```

Let's tie this exercise back to the issue raised in point 3. This simulation illustrated that quantifying the mediated effect proves difficult when background predictive variables confound the relationship between M and Y. Because M is not randomly assigned, it is important for us to think about how likely it is that our M and our Y are both affected by unobserved variables. In principle, if there are no confounding variables in this relationship, then a decomposition of effects analysis may be unbiased, but this is assumption is strong and usually hard to prove.

While we did not demonstrate it in this simulation, it is also possible to show that decomposition of effects also breaks down when treatment effects are heterogeneous (we introduced the intuition for this in point 4). The technical reason for this comes from our law of expectations, which is: $E[a*b]E[a*b] = E[a]E[b]+cov(a,b)E[a]E[b]+cov(a,b)$. If we have constant treatment effects, then aa and bb do not covary, the covariance term drops out, and we can simply multiply $a*ba*b$ to get the ACME. However, if the covariance term is non-zero, then we are not able to estimate this indirect effect from these two coefficients obtained from separate regressions. We constructed constant treatment effects in order to be able to demonstrate the process of decomposing effects, but if we were to re-do the simulation with heterogeneous treatment effects that covary, then we would not even be able to calculate the ACME or ACDE using the potential outcomes approach at the beginning of our code.

What you can do… Before you embark on a decomposition of effects analysis, ask yourself:

- Can I imagine any unobserved variables that predict both M and Y?
- Is it possible that my subjects respond to treatment effects in different ways?

If the answer to any of these questions is yes, we strongly recommend that you proceed with caution. In particular, think carefully about how unobserved variables and heterogeneous treatment effects would affect your estimation strategy.

7 Sometimes subgroup analysis can provide suggestive evidence for or against a mechanism.

In points 3-6, we've cautioned researchers against trying to confidently quantify the proportion of an effect that is mediated by a particular mechanism, but there may be other ways to learn more about mechanisms at work in a particular study. In point 1, we underscored the tight relationship between mechanisms and theory. Just because it's challenging to quantify evidence of a mechanism directly does not mean we cannot explore the testable predictions of the theory in which our mechanism is featured!

One strategy is to use subgroup analysis, or treatment-by-covariate interactions, to see whether different populations respond to the treatment differently in accordance with our theories. For example, suppose we wanted to learn more about the role of income in mediating the climate/conflict relationship. One of the testable implications of a theory in which income plays a mediating role is that we would expect climate shocks to be associated with conflict in areas where income is sensitive to climate shocks but not where income is independent of climate shocks. Sarsons (2015)[7] does exactly this. Exploiting the fact that districts downstream of irrigation dams do not depend on rainfall for income while districts upstream of irrigation dams do, she explores the income mechanism by testing whether rain shocks predict riot incidence in downstream districts but not in upstream districts. Formally, she tests these hypotheses:

- X→→Y in places where X is known to affect M [Rainfall shocks will increase riots in areas where rainfall shocks will negatively affect income (upstream of dam).]
- X has no effect on Y in places where X has no effect on M [Rainfall shocks will have no effect on riots in areas where income is not sensitive to rainfall (downstream of dam).]

However, she found that the relationship between rainfall shocks and riot incidence held just as tightly in the downstream districts where income was not sensitive to rainfall. She interprets this finding as "suggestive" evidence *against* the income mechanism. To be clear, Sarsons did not conduct any mediation analysis: she did not measure the income of each village and quantify the direct effect of rainfall shocks on riots and the indirect effect of rainfall shocks on riots through income. Instead, she looked for the heterogeneous treatment effects the theory would have implied and, finding no evidence of them, concluded that the income channel may be less important than previously thought.

What you can do… In future projects, ask yourself: If the mechanism is M, which groups or units would I expect to exhibit a treatment effect, and which groups or units would I expect not to respond to treatment? Next, test whether these predictions are supported by your data and interpret this as suggestive evidence for or against your proposed mechanism M. Keep in mind that such evidence is not decisive because the groups could differ in other ways that could affect their responsiveness to the treatment.

8 We can also look for suggestive evidence by looking at the effects of our treatment on various outcomes.

Again, while it's difficult to quantify evidence of a mechanism at work, we can always explore the testable implications of the theory in which our mechanism features. In point 7, we did this by exploring whether the treatment affected the particular subgroups for which a treatment effect is implied by our theory. Another approach is to explore whether the treatment affects only the outcomes implied by our theory.

For example, many social scientists are interested in how mass education influences democracy. Several theories of democratization expect different mechanisms would connect education and democracy. First, according to modernization theory, education could facilitate the smooth functioning of democracy by undermining group attachments (such as ethnicity or religion) in favor of merit.[8] Second, according to social theorists of oppression, education could undermine democracy by reinforcing obedience to authority, which is inherent in a classroom structure.[9] Third, according to many political scientists and psychologists, education can encourage democratic participation by empowering individuals with the ability to acquire and act on knowledge.[10] Friedman et al. (2011)[11] decide to tease apart these mechanisms by investigating the results of a field experiment in which Kenyan girls were randomly assigned to receive an education subsidy. They followed up with the students five years after the program and asked them several questions designed to test which of these three mechanisms were at work: Did the girls accept a husband's right to beat his wife? Was a parent involved in selecting their spouse? How strongly did the girl identify with her religious or ethnic group? Did the girl regularly read news?

The following table outlines the direction of the effects that each theory would suggest. Note that the various mechanisms being tested here result from theories with diverging predictions on some of these outcomes. The predictions from each of the three mechanisms are outlined on the rows, followed by the actual results. We can see that two of the outcomes collected provided support for modernization theory. However, modernization theory would have predicted a decrease in religious or ethnic group association (in reality, there was no effect) and had no predictions for newspaper readership (in reality, readership increased). None of the predictions of the obedience to authority mechanism were supported by the data. However, the data supported all four of the predictions of individual empowerment theory. The authors conclude that it is more likely
that $X \rightarrow M3 \rightarrow Y$ $X \rightarrow M3 \rightarrow Y$ than $X \rightarrow M1, M2 \rightarrow Y$ $X \rightarrow M1, M2 \rightarrow Y$.[12]

Mechanism	Acceptance of husband's right to beat wives (Y1)	Parent involved in selecting spouse (Y2)	Association with religion, ethnic identity (Y3)	Reads news (Y4)
(M1) Modernization	↓↓	↓↓	↓↓	No effect

Mechanism	Acceptance of husband's right to beat wives (Y1)	Parent involved in selecting spouse (Y2)	Association with religion, ethnic identity (Y3)	Reads news (Y4)
(M2) Obedience to authority	↑↑	↑↑	↑↑	No effect
(M3) Individual empowerment	↓↓	↓↓	No effect	↑↑
Actual effect	↓↓	↓↓	No effect	↑↑

This study, like Sarsons's study, is not trying to quantify how much of the effect of X on Y is conveyed via M. However, through thoughtful investigation of various outcomes, the authors are able to provide suggestive evidence of which mechanisms seem most plausible.

What you can do… In future projects, ask yourself: If the mechanism is M, which outcomes would I expect to be affected by my treatment, and which outcomes would I expect not to be affected by my treatment? Next, test whether these predictions are supported by your data and interpret this as suggestive evidence for or against your proposed mechanism M.

9 Designing complex treatments can help narrow our understanding of what part of the treatment is "doing the work."

Sometimes experimental researchers will try to better understand mechanisms by adding or subtracting elements of the treatment that are thought to trigger different mechanisms. This approach is sometimes called "implicit mediation analysis" because different components are X are thought to *implicitly* manipulate certain mechanisms. This, of course, is an assumption: because we are not measuring M directly, we are relying on a theoretical claim that component A will trigger M, whereas component B will not.

For example, many governments including Mexico, Brazil, Tanzania, and Uganda have created conditional cash transfer programs to address poverty. These programs provide cash to poor individuals, but they frequently come with conditions such as attending school or a job training program. Until recently, we knew only that these programs (X) successfully reduced poverty (Y) and that X caused Y either via cash or via the required attendance at school or job programs. To distinguish between these mechanisms, Baird et

al. (2011)[13] conducted an experiment in Malawi, where they assigned one group of families to receive a *conditional* cash transfer for the regular school attendance of their girls, another group of families to receive the cash *unconditionally*, and a control group to receive no transfer. This design "implicitly" manipulated M: while girls in the unconditional transfer group could also seek out education, school attendance (the condition under study) would likely be higher in the group that was required to seek it out. Unsurprisingly, school attendance and test performance was better for the group receiving conditional cash transfers. However, their measures of Y–the rate at which the girls became pregnant or married–were actually better (lower) in the group receiving the unconditional cash transfers. The authors concluded that attendance requirements associated with conditional cash transfers were not probably not the mechanism responsible for the success of these programs in reducing the symptoms of poverty.

Studies like these help not only social scientists to learn more about the channels through which X causes Y, but also policymakers to explore and discover new treatments. After several other studies joined Baird et al. in demonstrating the remarkable effects of unconditional cash transfers, many governments and organizations have begun to implement unconditional cash transfer programs.

What you can do… In future projects, ask yourself: Can my treatment be "unpacked" into multiple treatment arms, some that implicitly manipulate M, and some that do not? Consider using a factorial design to identify the effects of different treatment arms. If you have ample power, comparing the various treatment arms will provide you with suggestive evidence for or against M.

10 Despite the difficulties in empirically measuring mechanisms, it is worth paying serious attention to them but being cautious in our language.

Attempting to identify causal mechanisms is a noble endeavor. Articulating causal mechanisms is what allows us to unpack "black box" treatments and understand why and how certain treatments work. Even though causal claims can be (and often are) made without evidence for a causal mechanism, exploring causal mechanisms is what enables us to extend the research frontier and re-evaluate how our evidence maps on to our theories. For these reasons, audiences (be they the general public or academic reviewers) are often understandably eager for you to expound upon causal mechanisms after demonstrating evidence for a provocative causal claim. In anticipation of this, it is worth considering whether it is possible to design a way to test causal mechanisms in advance of implementing an experiment. If not, consider whether certain outcome measures or treatment by covariate interactions would provide some support for a particular causal mechanism, and be explicit about the limitations of this kind of analysis in your write-up. Mechanisms are an exciting domain of inquiry and should be considered both in the design and analysis of an experiment, but we should be sure to discuss mechanisms with caution appropriate to our ability to identify a particular mechanism and avoid overselling the argument.

1. Solomon M. Hsiang, Kyle C. Meng, and Mark A. Cane, "Civil Conflicts Are Associated with the Global Climate," Nature 476.7361 (2011): 438-441.
2. Ted Gurr, Why Men Rebel, Princeton University Press, 1970.
3. Alberto Chong, Ana L. de la O, Dean Karlan, and Leonard Wantchekon, "Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification," The Journal of Politics 77.1 (2015): 55-71.
4. Note that Chong et al. test their argument using data at the precinct level, not the individual level, but we've adapted their argument to the individual level for ease of exposition.
5. For a more rigorous discussion of these fallacies, see Adam N. Glynn, "The Product and Difference Fallacies for Indirect Effects," American Journal of Political Science 56.1 (2012): 257-269.
6. Explanation adapted from Alan Gerber and Donald Green, Field Experiments, W.W. Norton and Company, 2012, chapter 10.
7. Heather Sarsons, "Rainfall and Conflict: A Cautionary Tale." Journal of Development Economics 115 (2015): 62-72.
8. Marion Joseph Levy, Modernization and the Structure of Societies, Princeton University Press, 1966.
9. Frantz Fanon, The Wretched of the Earth, Grove Press, 1964. John Lott, Jr., "Public Schooling, Indoctrination and Totalitarianism," Journal of Political Economy 107(6), 1999.
10. Gabriel Almond and Sidney Verba, The Civic Culture: Political Attitudes and Democracy in Five Nations, Sage Publications, 1963. Robert Mattes and Michael Bratton, "Learning about Democracy in Africa: Awareness, Performance, and Experience," American Journal of Political Science, 51(1), 2007.
11. Willa Friedman, Michael Kremer, Edward Miguel, and Rebecca Thornton, "Education as Liberation?" NBER Working Paper 16939, 2011.
12. In the actual study, the authors were surprised to uncover evidence that education also increased individuals' acceptance of political violence. While they still argue that individual empowerment is responsible for the relationship between education and democracy, they caution that education does not always lead to democratization (that is, M3→Y but it is also possible that M3→NOT Y). Nonetheless, their approach is a useful demonstration of how multiple outcomes may shed light on mechanisms.
13. Sarah Baird, Craig McIntosh, Berk Ozler, "Cash or Condition? Evidence from a Cash Transfer Experiment," Quarterly Journal of Economics 126, 2011.