Causal Mediation Analysis

Down the rabbit hole

Lucas Shen

November 13, 2023

Table of contents

1 Statement of need

Mediation analysis is not new (Judd and Kenny 1981). But it is to me. Mediation analysis aims to shed light on mechanisms between some exposure and some outcomes of interest. It's an ambitious aim but, if well identified, has potential implementational and policy implications. Funding agencies (e.g., National Institute of Mental Health, NIH) for research grants encourage or even require mediation analyses of interventions to understand mechanisms (Nguyen, Schmid, and Stuart 2021). "Mediation analysis" is also increasingly popular. Since it's increasingly popular and results may influence policy and practice, I thought I should at least gain some literacy on mediation analysis. This vignette collects the gist of what I've managed to get from the literature. Figure 1.1 (taken directly from Nguyen, Schmid, and Stuart 2021) should sum up this statement of need.

1.1 (Un-mediated) Treatment effects

First, to set some context with the causal inference approach with (model-agnostic) potential outcomes (`PO`) without mediation. In a simple setup with just treatment T and an outcome of interest Y (without the mediator M for now), the main estimand of interest is the average treatment effect (`ATE`) defined as

$$\tau \equiv \mathbb{E}\big[Y(1)\big] - \mathbb{E}\big[Y(0)\big], \tag{1.1}$$

for the simplest scenario where T is a binary split:

$$T = \begin{cases} 1 & \text{if treated,} \\ 0 & \text{if untreated.} \end{cases}$$

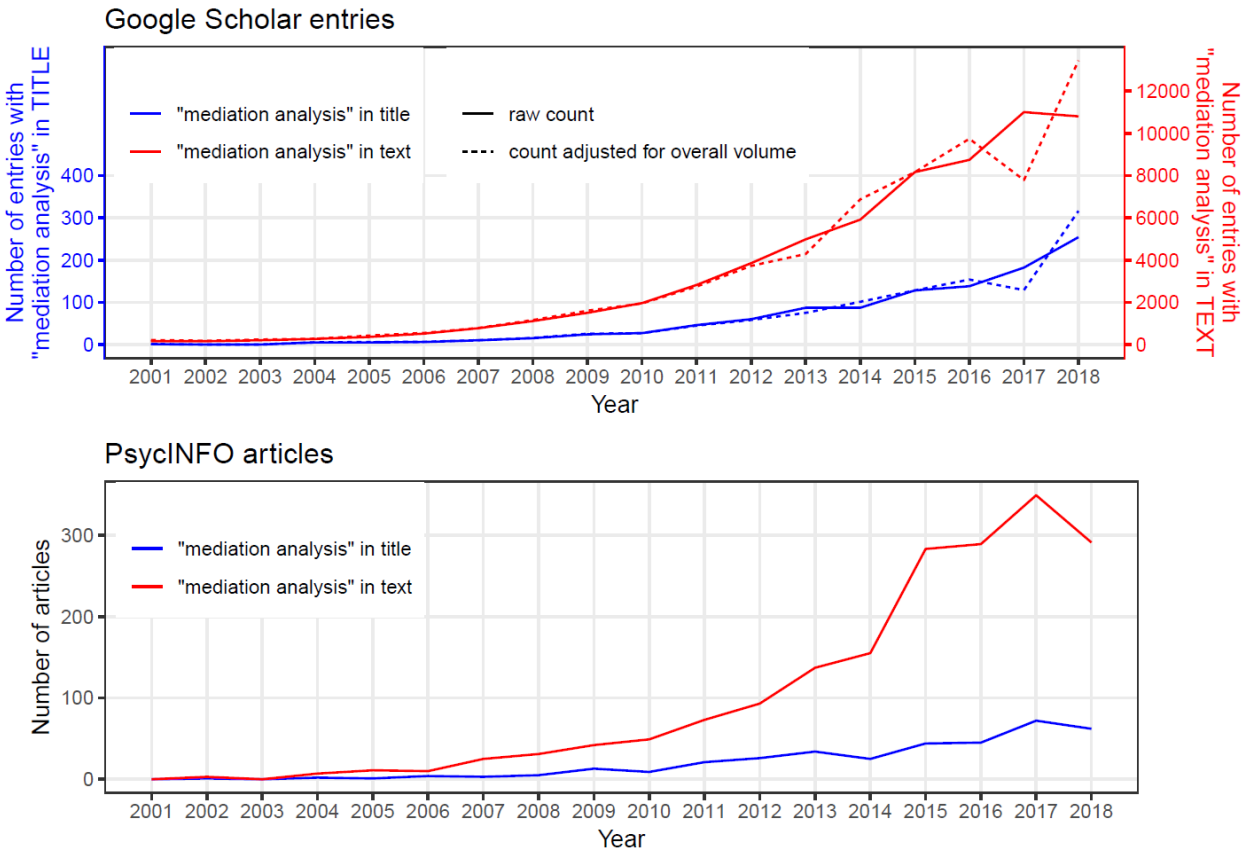
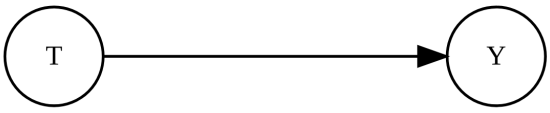


Figure 1.1: The raw counts in the top panel are counts reported by Google Scholar on two searches for articles (excluding patents and citations) with "mediation analysis" in the title, and for those with the same phrase anywhere in the text; the adjusted counts are adjusted for the fact that the volume of all Google Scholar entries varies in size from year to year, using 2015 as the standard year. In the bottom panel, the counts are reported by PsycINFO on the same two searches. These searches were conducted on 20/12/2018. ***Source:*** Figure 1 of Nguyen, Schmid, and Stuart (2021). Reproduced from v1 with permission.

$Y(1)$ is the potential outcome with treatment ($T = 1$) and $Y(0)$ is the potential outcome without treatment ($T = 0$). For a given individual, only $Y(1)$ is observed with treatment. $Y(0)$ is the counterfactual.[1]



Graph for the typical estimand of interest for T and Y—the average treatment effect (`ATE`, Equation 1.1). Arrows imply causality.

In reality, all we observe are outcomes conditional on the treatment $[Y|T = t]$ and the *statistical solution* (Holland 1986) is via differences in averages

$$\mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$$

which reduces to Equation 1.1 under certain conditions such as the `independence assumption`:[2]

$$Y_i(t) \perp\!\!\!\perp T_i. \tag{1.2}$$

The simplest way to estimate τ, given the assumptions are met, is

$$Y_i = \alpha + \tau T_i + \varepsilon_i.$$

1.2 Mediation through the lens of potential outcomes

I was wondering how mediation analysis fits into this barebones potential outcomes framework. Most publications with mediation analysis are in the Baron and Kenny tradition [Judd and Kenny (1981); Baron and Kenny (1986); MacKinnon (2012); hereafter, the classical mediation approach]. Inherent in mediation analysis, whether made explicit or not, is a causal hypothesis (Nguyen, Schmid, and Stuart 2021). T causes Y through M. There is, however, nothing in the mechanics of the classical approach that guarantees that uni-direction.[3]

[1] Aka missing data as the ''Fundamental Problem of Causal Inference" (Holland 1986).

[2] The conditional analog is similar for some baseline set of observables $\mathbf{X}$: $[Y_i(t) \perp\!\!\!\perp T_i]\big|\mathbf{X}_i$ (e.g., Imbens 2004; Angrist and Pischke 2009). *This vignette will largely abstract away from* $\boldsymbol{X}$ *for simplicity.*

[3] A "mediation analysis" without this explicit path (partly implied by the temporal ordering) is really just a "third variable analysis", which is an analysis of association (Nguyen, Schmid, and Stuart 2021).

So mediation could benefit from, or deserves, explicit causal thinking (Nguyen, Schmid, and Stuart 2021).[4] Work to think of mediation in the causal inference framework started as early as Robins and Greenland (1992) and as recent (at time of writing) as Stuart et al. (2022) (hereafter, the causal inference approach).[5] With a proposed mediation path through a variable M, the analog of the `ATE` (from Equation 1.1) is now: [6]

$$\tau = \mathbb{E}\big[Y\left(1, M(1)\right)\big] - \mathbb{E}\big[Y(0, M(0))\big], \tag{1.3}$$

where (like T) the simplest scenario is:

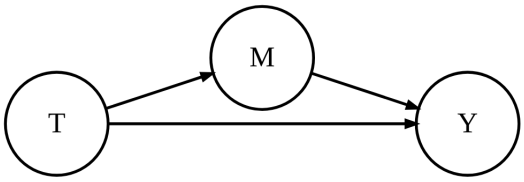
$$M = \begin{cases} 1 & \text{if T} = 1, \\ 0 & \text{if T} = 0, \end{cases}$$

where the value of the mediator M also depends on T.

From the total effect (Equation 1.3),[7] we can start decomposing into the direct and indirect effects.[8] Apparently, there are two different sets of estimands (Nguyen, Schmid, and Stuart 2021). No one told me that…

The direct effect is:

$$\zeta(t) \equiv \mathbb{E}\big[Y_i(1, M_i(t))\big] - \mathbb{E}\big[Y_i(0, M_i(t))\big],\ t \in \{0, 1\}. \tag{1.4}$$



Graph of mediation analysis. The path between T and Y is the direct effect (Equation 1.4). The mediation effect is captured from T to M and then M to Y (Equation 1.5).

[4] This is different from other regular regression analyses, where, depending on the target and assumptions, both causal and conditional association interpretations are valid (Nguyen, Schmid, and Stuart 2021).

[5] Separate issue of temporal ordering of T, M, and Y (not covered in this vignette. See P49 of Stuart et al 2021 for more references.)

[6] This PO notation is similar to that in Imbens (2004), Angrist and Pischke (2009), Kosuke Imai, Keele, and Tingley (2010); Nguyen, Schmid, and Stuart (2021), etc. Pearl (2001), Hernan and Robins (2023), etc., have slightly different PO notations.

[7] The *total effect* (`TE`) is used in place of the `ATE` in the language of mediation analysis (e.g., Judd and Kenny 1981; Robins and Greenland 1992; K. Imai, Keele, and Yamamoto 2010; Bullock, Green, and Ha 2010; Nguyen, Schmid, and Stuart 2021; Hernan and Robins 2023).

[8] While going through a bunch of resources from different fields, I came across a suggestion that the direct effect should really be called the unmediated effect (although I was unable to trace back the source). The notion is that the *direct effect* is no more a direct effect than the residual is a Gaussian error. If there are unspecified mediation pathway(s), the direct effect will subsume those.

The indirect effect is:

$$\delta(t) \equiv \mathbb{E}\big[Y_i(t, M_i(1))\big] - \mathbb{E}\big[Y_i(t, M_i(0))\big],\ t \in \{0, 1\}. \tag{1.5}$$

Varying t gives two different sets of estimands for each of δ and ζ.

1.3 Set A: ADTE and AITE

The first set of estimands for direct and indirect effects are the average direct treatment effect

$$ADTE \equiv \zeta(0) = \underbrace{\mathbb{E}\big[Y(1, M(0))\big] - \mathbb{E}\big[Y(0, M(0))\big]}_{\text{effect of } T \text{ on } Y \text{ when } M \text{ is held at M(T=0)}} \tag{1.6}$$

and the average indirect treatment effect.

$$AITE \equiv \delta(1) = \underbrace{\mathbb{E}\big[Y(1, M(1))\big] - \mathbb{E}\big[Y(1, M(0))\big]}_{\text{effect of } M \text{ on } Y \text{ when } T = 1} \tag{1.7}$$

The decomposition of the `ATE` into the two effects follows directly from Equation 1.3:

$$\underbrace{\mathbb{E}\big[Y(1, M(1))\big] - \mathbb{E}\big[Y(0, M(0))\big]}_{\text{total effect}} = \mathbb{E}\big[Y(1, M(1))\big] \underbrace{- \mathbb{E}\big[Y(1, M(0))\big] + \mathbb{E}\big[Y(1, M(0))\big]}_{\text{additive identity}} - \mathbb{E}\big[Y(0, M(0))\big] \tag{1.8}$$

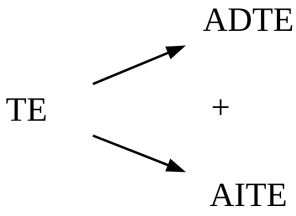
$$= \underbrace{\mathbb{E}\big[Y(1, M(1))\big] - \mathbb{E}\big[Y(1, M(0))\big]}_{\text{AITE}} + \underbrace{\mathbb{E}\big[Y(1, M(0))\big] - \mathbb{E}\big[Y(0, M(0))\big]}_{\text{ADTE}} \tag{1.9}$$

The `AITE`, in particular, is also known as the causal mediation effect.[9] This decomposition is suitable when there are strong

[9] Sometimes `ACME` (average causal mediation effect, K. Imai, Keele, and Yamamoto 2010; Kosuke Imai, Keele, and Tingley 2010). ADTE is aka the natural direct effect (NDE). AITE is aka the natural indirect effect (NIE). That's Pearl's language (Pearl 2001). Another one is pure direct effect and total indirect effect (e.g., Robins and Greenland 1992; Hernan and Robins 2023).

priors about a direct effect but unclear priors on whether an indirect effect also exists. An example could be where T is a STEM degree that raises labor market value Y through some technical competency. On the other hand, it is unclear whether additional benefits also arise through networking with peers M.

Another example could be where T is a dietary program (e.g., a low-whatever-the-frack-is-the-new-fad diet) that lowers diabetic risk Y, but it is unclear if the reduction in diabetic risk also comes from weight loss M. If not, targeting weight loss may be of limited efficacy. This is of practical implication since it informs where and what type of intervention should occur.



Set A: ADTE and AITE. ADTE is Equation 1.4 with $t = 0$—holding $m = 0$ and asking what is the direct effect of with T. AITE is Equation 1.5 with $t = 1$—holding $t = 1$ and asking what is the indirect effect with M.

1.4 Set B: ADTET and AITEC

The second set of estimands that decompose the `ATE` are the average direct treatment effect wrt the treated

$$ADTET \equiv \zeta(1) = \underbrace{\mathbb{E}[Y(1, M(1))] - \mathbb{E}[Y(0, M(1))]}_{\text{effect of T when } M \text{ is held at M(T=1)}} \quad (1.10)$$

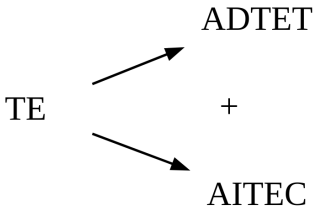
and the average indirect treatment effects wrt to controls

$$AITEC \equiv \delta(0) = \underbrace{\mathbb{E}[Y(0, M(1))] - \mathbb{E}[Y(0, M(0))]}_{\text{effect of M when } T=0}. \quad (1.11)$$

Decompositon again follows directly:

$$\underbrace{\mathbb{E}[Y(1, M(1))] - \mathbb{E}[Y(0, M(0))]}_{\text{total effect}} = \underbrace{\mathbb{E}[Y(1, M(1))] - \mathbb{E}[Y(0, M(1))]}_{\text{ADTET}} + \underbrace{\mathbb{E}[Y(0, M(1))] - \mathbb{E}[Y(0, M(0))]}_{\text{AITEC}}.$$

This decomposition is useful when there are strong priors about an indirect mediating path but unclear priors about a direct effect. An example could be where T is an MBA degree with a weak connection labor market value Y through some technical competency but where benefits mainly arise through networking in the MBA program M.



Set B: ADTET and AITEC. ADTET is Equation 1.4 with $m = 1$—holding $m = 1$ and asking what is the direct effect of with T. AITEC is Equation 1.5 with $t = 0$—holding $t = 0$ and asking what is the indirect effect with M.

1.5 Classical mediation approach

In the classical mediation approach, in addition to

$$Y_i = \alpha + \beta T_i + \varepsilon_i,$$

two equations are estimated in a structural equation model(Baron and Kenny 1986). One modeling T and its effect on M. The other modeling the effect of T and M on Y.

$$M_i = \alpha_0 + \alpha_1 T_i + \varepsilon_{1i} \tag{1.12}$$

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 M_i + \varepsilon_{2i} \tag{1.13}$$

Equation 1.12 and Equation 1.13 are estimated separately. The mediation effect is observed via the "product-of-coefficients" method as $(\hat{\alpha_1}\hat{\beta_2})$—the effect of T on M multiplied by the effect of M on Y. The direct effect of T on Y would be observed with $\hat{\beta}_1$.

The conditional analog of this with a set of covariates $\mathbf{X}$ should follow closely (Judd and Kenny 1981).

1.6 Classical mediation approach: Common grounds

How would the causal inference approach with potential outcomes connect to the classical approach?

The reduced form (from Equation 1.12 and Equation 1.13) is:

$$\mathbb{E}[Y_i|T_i, M_i] = \beta_0 + \beta_1 \mathbb{E}[T_i] + \beta_2 \mathbb{E}[M_i]. \tag{1.14}$$

Together with the definition of the `AITE` (Equation 1.7), this gives

$$
\begin{aligned}
AITE &= \mathbb{E}[Y_i(1, M_i(1))] - \mathbb{E}[Y_i(1, M_i(0))] && (1.15)\\
&= \Big\{\beta_0 + \beta_1(1) + \beta_2(\alpha_0 + \alpha_1(1)\Big\} - \Big\{\beta_0 + \beta_1(1) + \beta_2(\alpha_0)\Big\} && (1.16)\\
&= \Big\{\alpha_0\beta_2 + \alpha_1\beta_2\Big\} - \Big\{\alpha_0\beta_2\Big\} && (1.17)\\
&= \alpha_1\beta_2 && (1.18)
\end{aligned}
$$

which is exactly the same mediation effect proposed by the "product-of-coefficients" from the structural equations above (Section 1.5).[10]

If the classical approach and the causal inference approach both agree on the underlying intuition of mediation, why bother with the potential outcomes framework? One reason is that it helps define estimands (as in Equation 1.4 in Section 1.2). Another key reason is that it helps clarify what the identification assumptions are.

1.7 Biases, bounds, and assumptions

What we only observe in the real world is $[Y_i|T = t, M = m]$, $t \in \{0, 1\}$, $m \in \{0, 1\}$. We also only observe actualized potential outcomes $Y(t, m)$ for the corresponding m and t values. The observed difference only converges to $\delta(t)$ (`AITE`) under certain conditions. The observed difference in a mediated effect, given that treatment is "`switched on`" $t = 1$, is

$$
\begin{aligned}
&\mathbb{E}[Y|T = 1, M = 1] - \mathbb{E}[Y|T = 1, M = 0]\\
&= \mathbb{E}[Y(1, 1)|T = 1, M = 1] - \mathbb{E}[Y(1, 0)|T = 1, M = 0] && (1.19)\\
&= \mathbb{E}[Y(1, 1)|M = 1] - \mathbb{E}[Y(1, 0)|M = 0] \qquad (\text{from } Y_i(t, m) \perp\!\!\!\perp T_i)
\end{aligned}
$$

where, even given randomization of T (analogous to Equation 1.2 in the un-mediated case), the selection issue with M

[10] Section A.2 in the Supplementary appendix shows that what falls out from the other three estimands is also consistent with the product of coefficients.

persists. It will carry its own selection bias unless additional assumptions are imposed. The above only reduces to the `AITE` by imposing additional assumptions on the statistical independence of M,

$$\begin{aligned}
&\mathbb{E}[Y(1,1)|M=1] - \mathbb{E}[Y(1,0)|M=0] \\
&= \mathbb{E}[Y(1,1)|M=1] - \underbrace{\mathbb{E}[Y(1,0)|M=1] + \mathbb{E}[Y(1,0)|M=1]}_{\text{additive identity}} - \mathbb{E}[Y(1,0)|M=0] \\
&= \underbrace{\mathbb{E}[Y(1,1) - Y(1,0)\Big|M=1]}_{\text{analog to AITE}} + \underbrace{\mathbb{E}[Y(1,0)|M=1] - \mathbb{E}[Y(1,0)|M=0]}_{\text{selection into M}}
\end{aligned} \tag{1.20}$$

where $E[Y(1,0)|M=1]$ in the first equality is a counterfactual—what would potential outcome be for people with the M `switched off` $m=0$ given that it is actually `switched on` $m=1$ (*concrete example* later below). The second equality collects terms into the underlying mediation estimand and an **additional bias component**.[11]

There is a tendency for this bias to be positive,[12] and, therefore, overstate the mediation effect. Take, as a *concrete example*, a case where we are interested in child income Y, parent income T, and college attendance M in a classic case of intergenerational income mobility. Parent income affects child income, but also potentially through access to better education. The expected sign of the AITE should be positive. The first part of the bias is the counterfactual of the potential income of the child of not going to college, given that she did. Conditional on parent income, the child that goes to college should have qualities (non-existent in the child that never goes to college) that would be helpful for labor market success, and, to this extent, suggest that the bias component is > 0. Hence, without intervening in the assignment of M, the estimated mediation effect tends to be biased. But the observed effect may still be useful to provide optimistic **bounds** for the underlying mediation effect.[13]

[11] Section A.3 in the Supplementary appendix offers an alternative algebraic insight of the bias.

[12] Or, generally, share the same sign as the AITE.

[13] This *bound* is *not* a statistical property. Instead, the bound is based on counterfactual reasoning about what the plausible potential outcomes are from domain understanding.

The last line in Equation 1.20 only reduces to the true `AITE` by imposing additional **assumptions**.

$$\left\{Y_i(t,m), M_i(t')\right\} \perp\!\!\!\perp T_i \tag{1.21}$$

$$\left[Y_i(t',m) \perp\!\!\!\perp M_i(t)\right] \Big| T_i = t \tag{1.22}$$

First, for a causal interpretation, the treatment assignment T has to be statistically orthogonal to the potential outcomes and the potential mediators. This should usually be satisfied with randomization of T as in the usual unmediated case (Section 1.1). This assumption was imposed in the last equality of Equation 1.19. Second, given the actual value $T_i = t$, the mediator M should be statistically orthogonal to the potential outcome. The standard proposed solution to achieve this is to subject M to experimental manipulation, just as with T.[14] The above is the `sequential ignorability assumption` since the assumptions are made *sequentially* (K. Imai, Keele, and Yamamoto 2010).

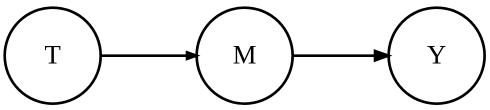
In addition, there is a (slightly more subtle) assumption with the common data support for T (and M).

$$\exists \theta > 0 : \ \theta \leq \mathbb{E}\left[T_i = t\right] \leq (1-\theta), \qquad t \in \{0,1\}$$

With mediation, an additional assumption applies to the mediator:

$$\exists \eta > 0 : \ \eta \leq \mathbb{E}\left[M_i = m \middle| T_i = t\right] \leq (1-\eta), \qquad t \in \{0,1\}.$$

Loosely speaking, with mediation, twice the identification assumptions are needed.



Graph of mediation but ruling out a direct effect from T to Y. Aka the instrumental variables design where canonical assumptions have been established as early as (Angrist, Imbens, and Rubin 1996).

1.8 Mediation and instrumental variables

What if one thinks that the effect of T should only pass purely through M? In this case, the graph in Figure **??** reduces to

[14] Section A.4 in the Supplementary appendix shows an example of how the AITE cannot be identified even with randomization if heterogeneities in effects exist.

Figure ??. This turns out to be the causal graph of the instrumental variables approach. However, the catch now is that the set of identification assumptions changes, with the sequential ignorability (Equation 1.21, Equation 1.22) no longer relevant (K. Imai, Keele, and Yamamoto 2010). Instead, the identification assumptions are those canonically established in the instrumental variable framework, such as that of exclusion restriction (Angrist, Imbens, and Rubin 1996). Moreover, the instrumental variable framework by design a priori rules out other causal mechanisms. This may be less than ideal for many health and social science research (K. Imai, Keele, and Yamamoto 2010).

1.9 Mediation with moderation

Mediation is not moderation. Mediation allows the treatment T affect the outcome of interest Y through some proposed mediator M indirectly. Moderation, on the other hand, allows for the effect of T to differ by levels of the moderator (e.g., vaccine efficacy differing by age groups).[15] But it turns out the mediation analysis framework also supports moderation with an additional moderating term $(M_i \times T_i)$ in Equation 1.13.

$$\begin{aligned} M_i &= \alpha_0 + \alpha_1 T_i + \varepsilon_{1i} \\ Y_i &= \beta_0 + \beta_1 T_i + \beta_2 M_i + \beta_3 M_i T_i + \varepsilon_{2i} \end{aligned} \tag{1.23}$$

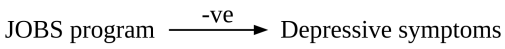
[15] Regarding temporal ordering, moderators should be pre-exposure baselines (related to colliders/bad controls not covered here), while mediators should be post-exposure. Analyses termed as "mediation analyses", but where it is impossible for T to influence M because of temporal ordering, are really "third variable analyses" (Nguyen, Schmid, and Stuart 2021). Section A.1 shows a slight trend in "causal mediation analysis" only very recently.

Again, using the definition of the `AITE` (Equation 1.7)

$$AITE \equiv \delta(1) \tag{1.24}$$
$$= \mathbb{E}[Y_i(1, M_i(1))] - \mathbb{E}[Y_i(1, M_i(0))] \tag{1.25}$$
$$= \Big\{\beta_0 + \beta_1 + \beta_2(\alpha_0 + \alpha_1) + \beta_3(\alpha_0 + \alpha_1)\Big\} - \Big\{\beta_0 + \beta_1 + \beta_2(\alpha_0) + \beta_3(\alpha_0)\Big\} \tag{1.26}$$
$$= \beta_2\alpha_1 + \beta_3\alpha_1 \tag{1.27}$$
$$= \alpha_1(\beta_2 + \beta_3) \tag{1.28}$$

which is just the effect of T on M (the α_1) multiplied by the coefficients of M in the outcome regression (β_2 and β_3).

So, the mediation framework also directly allows for the mediator to have a role as a moderator. Compared to just $\alpha_1\beta_2$ from Section 1.6, the additional $\alpha_1\beta_3$ captures the additional effect, if any, of the mediator M for those individuals in the treatment group $T = 1$. $\alpha_1\beta_3$ is also another way of describing how M moderates the effect of the treatment T. If it turns out $\beta_3 = 0$, then this implies no moderating relationship exists, and the model reverts to that in Equation 1.13.[16]



From past studies, participants randomized into treatment groups (`JOBS program`) have been found to have lower levels of depressive symptoms (Vinokur and Schul 1997).

1.10 Application example: JOBS II

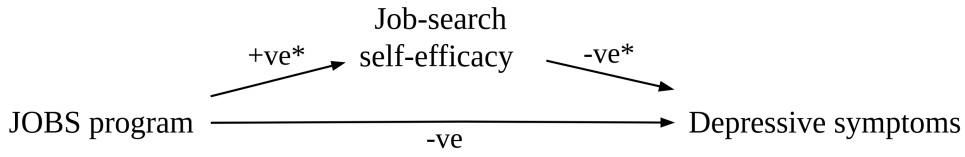
To give another concrete example, Kosuke Imai, Keele, and Tingley (2010) (a running reference in this note) use the JOBS II experiment from the psychology literature on (un)employment and mental health (e.g., Vinokur and Schul (1997)). JOBS (Job Opportunities and Basic Skills training) II is a randomized field experiment. The goal was to identify the effects of a job training intervention on employment status and also on the mental wellbeing of the job seekers.[17]

[16] See Section A.5 in the Supplementary appendix for the `ADTE` with moderation.

[17] JOBS II had 1,801 participants randomized into treatment and control groups. The treatment group participated in job skills workshops that taught job search skills and coping mechanisms. I.e., they got the `JOBS program`. For mental wellbeing, the key outcome measure is a continuous measure of `Depressive symptoms` using the Hopkins Symptom Checklist (subscale of 11 items).

An extension of this interest is also to understand mediator factors. Past studies found the program helpful in reducing depressive symptoms (Figure **??**). If true, what would be a mediating factor? Credibly identifying mediator factor(s) has implications for future programs. Furthermore, would the mediating factor be moderated by treatment status? This would help identify high-risk subgroups that can be more effectively screened in future programs (Vinokur and Schul 1997).

The mediator that is evaluated is `Job-search self-efficacy`.[18] Importantly, self-efficacy is not randomized, so the second ignorability assumption is not satisfied by design. Instead, baseline covariates measured before the workshops (e.g., education, income, race, marital status, age, sex, previous occupation, and the level of economic hardship) are adjusted for. It turns out that a mediation effect exists. The estimated mediation effect is negative, which implies that JOBS helped reduce depressive symptoms by increasing job-search self-efficacy. The direct effect and the total effects, however, have null estimated effects. Moderating the mediator also yields null estimated effects. Figure 1.10 summarizes the relationship.



Job-search self-efficacy as a mediating effect: Participants with higher self-efficacy report lower levels of depressive symptoms. The $*$ denotes estimates found to be statistically distinguishable from zero at conventional levels.

[18] Self-efficacy comes from six (five-point scale) items on the degree of confidence in being able to successfully perform six essential job-search activities (e.g., completing job application/resume, using social network to discover promising job openings, getting point across in a job interview; Vinokur and Schul (1997)).

1.11 Endnote

No new grounds have been broken by me here. This vignette is merely my attempt at parsing and then synthesizing different resources that already exist (some as recent as Stuart et al. (2022) in the Journal of Causal Inference) to gain some literacy on (causal) mediation analysis. A lot of studies use mediation to imply some directional effect. This clearly alludes to a causal channel. Yet it's rarely clear what the estimands are and what identification assumptions are needed. This tracks with the exegesis by Bullock, Green, and Ha (2010) and Nguyen, Schmid, and Stuart (2021). What's covered here should make identification assumptions more transparent. If there's some error you want to point out or have comments, feel free to reach out to me. Also, not everything is covered here, but the list of canonical resources should lead to more details.

1.12 Resources

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *J. Am. Stat. Assoc.* 91 (434): 444–55. https://doi.org/10.2307/2291629.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Baron, Reuben M, and David A Kenny. 1986. "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *J. Pers. Soc. Psychol.* 51 (6): 1173–82. https://doi.org/10.1037/0022-3514.51.6.1173.

Bullock, John G, Donald P Green, and Shang E Ha. 2010. "Yes, but What's the Mechanism? (Don't Expect an Easy Answer)." *J. Pers. Soc. Psychol.* 98 (4): 550–58. https://doi.org/10.1037/a0018933.

Hernan, Miguel A, and James M Robins. 2023. *Causal Inference*. CRC Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *J. Am. Stat. Assoc.* 81 (396): 945–60. https://doi.org/10.2307/2289064.

Imai, K, L Keele, and T Yamamoto. 2010. "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects." https://projecteuclid.org/journals/statistical-science/volume-25/issue-1/Identification-Inference-and-Sensitivity-Analysis-for-Causal-Mediation-Effects/10.1214/10-STS321.short.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychol. Methods* 15 (4): 309–34. https://doi.org/10.1037/a0020761.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Rev. Econ. Stat.* 86 (1): 4–29. https://doi.org/10.1162/003465304323023651.

Judd, Charles M, and David A Kenny. 1981. "Process Analysis: Estimating Mediation in Treatment Evaluations." *Eval. Rev.* 5 (5): 602–19. https://doi.org/10.1177/0193841X8100500502.

MacKinnon, D. 2012. "Introduction to Statistical Mediation Analysis."

Nguyen, Trang Quynh, Ian Schmid, and Elizabeth A Stuart. 2021. "Clarifying Causal Mediation Analysis for the Applied Researcher: Defining Effects Based on What We Want to Learn." *Psychol. Methods* 26 (2): 255–71. https://doi.org/10.1037/met0000299.

Nguyen, Trang Quynh, Ian Schmid, and Elizabeth A. Stuart. 2020. "Clarifying Causal Mediation Analysis for the Applied Researcher: Defining Effects Based on What We Want to Learn." https://arxiv.org/abs/1904.08515.

Pearl, Judea. 2001. "Direct and Indirect Effects." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–20. UAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. https://dl.acm.org/doi/10.5555/2074022.2074073.

Robins, James M, and Sander Greenland. 1992. "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology* 3 (2): 143–55. http://www.jstor.org/stable/3702894.

StataCorp. 2023. "Mediate — Causal Mediation Analysis." *Stata Manuals*.

Stuart, Elizabeth A, Elizabeth L Ogburn, Ian Schmid, and

Trang Quynh Nguyen. 2022. "Clarifying Causal Mediation Analysis: Effect Identification via Three Assumptions and Five Potential Outcomes." *Journal of Causal Inference* 10 (1): 246–79. https://doi.org/10.1515/jci-2021-0049.

Vinokur, Amiram D, and Yaacov Schul. 1997. "Mastery and Inoculation Against Setbacks as Active Ingredients in the JOBS Intervention for the Unemployed." *J. Consult. Clin. Psychol.* 65 (5): 867–77. https://doi.org/10.1037/0022-006X.65.5.867.

A Supplementary appendix

A.1 Trends in mediation analysis

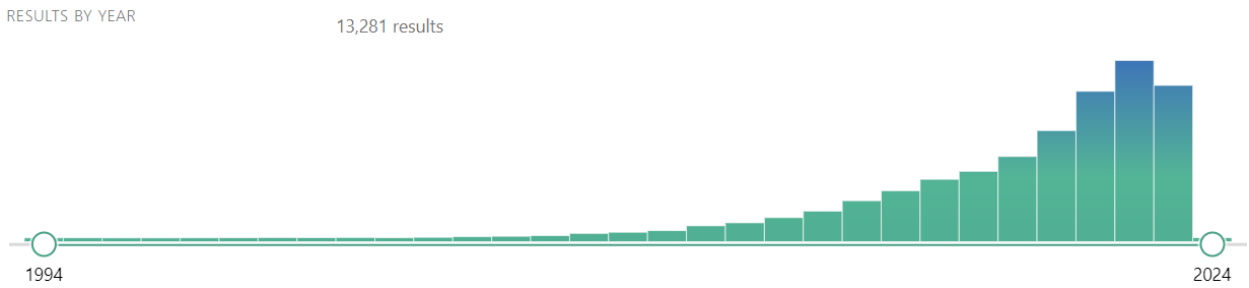


Figure A1: Trends in mediation analysis. Screenshot of the results from searching PubMed from database inception to end of 2023 for ("mediation analysis") OR ("mediation analyses") without any other filters. 13,281 results found, which has been trending up since around 2010. See Figure 1.1 for similar trends using Google Scholar and PsycINFO from Nguyen, Schmid, and Stuart (2021).

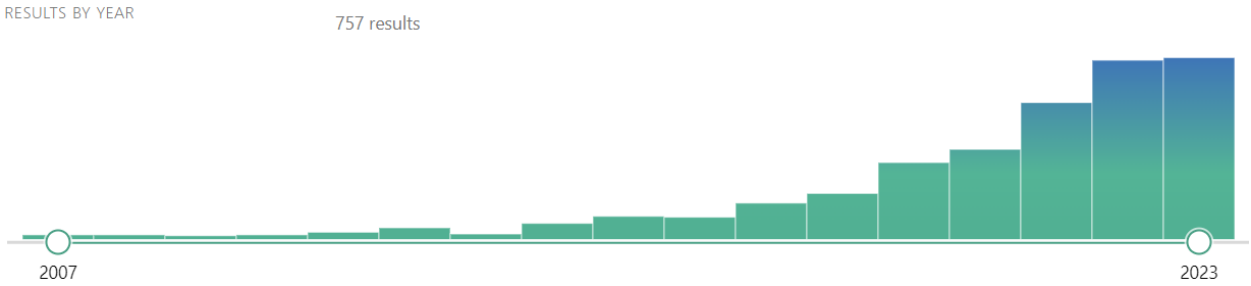


Figure A2: Trends in mediation analysis with explicit reference to causality. Screenshot of the results from searching PubMed from database inception to end of 2023 for ("causal mediation analysis") OR ("causal mediation analyses") without any other filters. 757 results found, which has been trending up since around 2018.

A.2 Classical mediation approach: ADTE, ADTET, AITEC

Together with the definition of the ADTE (Equation 1.7), this gives

$$
\begin{aligned}
ADTE = \zeta(0) &= \mathbb{E}[Y_i(1, M_i(0))] - \mathbb{E}[Y_i(0, M_i(0))] && (1.29)\\
&= \Big\{\beta_0 + \beta_1(1) + \beta_2(\alpha_0)\Big\} - \Big\{\beta_0 + \beta_1(0) + \beta_2(\alpha_0)\Big\} && (1.30)\\
&= \beta_1, && (1.31)
\end{aligned}
$$

which is exactly the same β_1 coefficient in Equation 1.13 capturing the direct effect of T on Y.

If the M is `switched on` $m = 1$, we get the same result for ADTET (Equation 1.10):

$$
\begin{aligned}
ADTET = \zeta(1) &= \mathbb{E}[Y_i(1, M_i(1))] - \mathbb{E}[Y_i(0, M_i(1))] && (1.32)\\
&= \Big\{\beta_0 + \beta_1(1) + \beta_2(\alpha_0 + \alpha_1(1)\Big\} - \Big\{\beta_0 + \beta_1(0) + \beta_2(\alpha_0 + \alpha_1(1))\Big\} && (1.33)\\
&= \beta_1. && (1.34)
\end{aligned}
$$

The AITEC Equation 1.11 also reduces to the product-of-coefficients $\alpha_1\beta_2$:

$$AITEC = \delta(0) = \mathbb{E}[Y_i(0, M_i(1))] - \mathbb{E}[Y_i(0, M_i(0))] \tag{1.35}$$

$$= \Big\{\beta_0 + \beta_1(0) + \beta_2(\alpha_0 + \alpha_1(1)\Big\} - \Big\{\beta_0 + \beta_1(0) + \beta_2(\alpha_0)\Big\} \tag{1.36}$$

$$= \alpha_1\beta_2. \tag{1.37}$$

A.3 Biases, bounds, and assumptions: Algebraic version

Bullock, Green, and Ha (2010) offer algebraic perspective of the bias seen in Section 1.7. Starting with the linear structural equation model:

$$Y_i = \alpha_1 + \beta_1 T_i + \varepsilon_{1i} \tag{1.38}$$

$$M_i = \alpha_2 + \beta_2 T_i + \varepsilon_{2i} \tag{1.39}$$

$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \varepsilon_{3i} \tag{1.40}$$

$$\tag{1.41}$$

The probability limit of the estimated mediated effect $\hat{\gamma}$ (Bullock, Green, and Ha 2010) is

$$\underbrace{\gamma}_{\text{AITE}} + \underbrace{\frac{cov(\varepsilon_{2i}, \varepsilon_{3i})}{var(\varepsilon_{2i})}}_{\text{asymptotic bias}},$$

where $cov(\varepsilon_{2i}, \varepsilon_{3i})$ is the covariance of the error terms in the un-mediated model and the mediated model. So $\hat{\gamma}$ is only consistent if the covariance term reduces to 0, which is unlikely because whatever unmeasured X_i is in ε_{2i} should also be present in ε_{3i} (like unmeasured qualities of a good worker in Section 1.7). Moreover, the sign of $cov(\varepsilon_{2i}, \varepsilon_{3i})$ tends to coincide with the sign of γ, leading to the overstatement of the magnitude of the mediation effect γ, if any (as discussed in Section 1.7).

A.4 Heterogeneities in mediation effect

Not as much, it seems, is said about heterogeneities in mediation analysis. Bullock, Green, and Ha (2010) do a treatment of this. Heterogeneities, together with the product-of-coefficients method, can be problematic even when M is already subject to experimental intervention.

A similar model to Equation 1.12 and Equation 1.13 with heterogeneities in effect is

$$M_i = \alpha_1 + \beta_{1i} T_i + \varepsilon_{1i} \tag{1.42}$$
$$Y_i = \alpha_2 + \beta_{2i} T_i + \gamma_i M_i + \varepsilon_{2i} \tag{1.43}$$

To see why heterogeneity could be bad without simulations, suppose we know that the true effects for two groups, A and B (equal proportions), are as follow:

$$\beta_{1A} = 2 \tag{1.44}$$
$$\gamma_A = 2 \tag{1.45}$$

and

$$\beta_{1B} = -3 \tag{1.46}$$
$$\gamma_B = -1. \tag{1.47}$$

The mediation effect for group A is $\beta_{1A}\gamma_A = 4$ and for group B it is $\beta_{1B}\gamma_B = 3$. So the average of the two (given equal group proportion) is $\frac{4+3}{2} = 3.5$.

However, the conventional estimate will not give 3.5. This is because

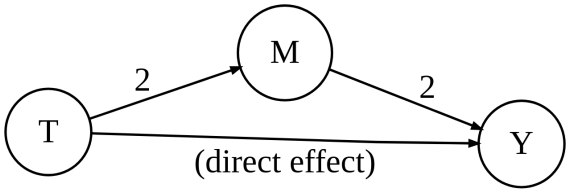
$$\bar{\beta}_i = \frac{\beta_{1A} + \beta_{1B}}{2} = \frac{2-3}{2} = -0.5$$

and

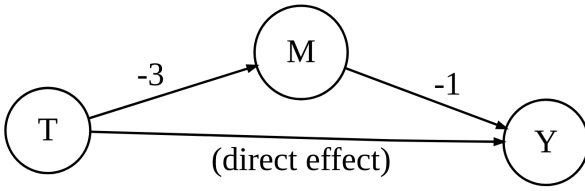
$$\bar{\gamma} = \frac{\gamma_A + \gamma_B}{2} = \frac{2-1}{2} = 0.5$$

The conventional estimate from a product of coefficients is -0.25 ($\neq 3.5$). This resembles basic probability, where

$$\mathbb{E}[\beta_{1i}]\mathbb{E}[\gamma_i] = \mathbb{E}[\beta_{1i}\gamma_i] - \underbrace{cov(\beta_{1i}, \gamma_i)}_{\neq 0}$$



Group A's indirect effect is $4(=(2)(2))$. The effect from T to M is 2, and the effect from M to Y is 2. Abstracting away from direct effects.



Group B's indirect effect is $3(=(-3)(-1))$. The effect from T to M is -3, and the effect from M to Y is -1. Abstracting away from direct effects.

so the product of the averages does not equal the average of the product unless that additional covariance term is zero. So even if M has been randomized, we still cannot identify the AITE (mediation effect) if heterogeneities exist (Bullock, Green, and Ha 2010).

A.5 Mediation with moderation: Direct effect

With moderation (as specified in Equation 1.23), and from the definition of the `ADTE` ($\zeta(1)$ from Equation 1.4):

$$ADTE \equiv \zeta(1) \tag{1.48}$$
$$= \mathbb{E}[Y_i(1, M_i(0))] - \mathbb{E}[Y_i(0, M_i(0))] \tag{1.49}$$
$$= \Big\{\beta_0 + \beta_1 + \beta_2(\mathbb{E}[M|T=0]) + \beta_3(\mathbb{E}[M|T=0])\Big\} - \Big\{\beta_0 + \beta_2(\mathbb{E}[M|T=0])\Big\} \tag{1.50}$$
$$= \beta_1 + \alpha_0\beta_3. \tag{1.51}$$

So the `ADTE` is the effect of T on Y and an additional effect that is the product of a constant for M and its effect moderated with T. If no moderation exists (i.e., $\beta_3 = 0$), the above reduces to just β_1, as in Section A.2. In other words, mediation with moderation (allowing for the interaction of the mediator and treatment status) is the general case. No interactions is the special case.