STATISTICS · CAUSAL INFERENCE

Causal Mediation Analysis for Decomposing Pathways of Knowledge Influence

How causal mediation analysis splits an effect into the part that works through knowledge or another intermediate variable and the part that does not, which assumptions that needs, and how to report it.

ModQu research article · 1,882 words · 19 references (APA 7th edition)

Knowing that something works is often not enough. A software company that finds its new onboarding tutorial raises customer retention will want to know *why*. Does the tutorial work because users learn more about the product, or because it creates a good first impression that has little to do with what they learn? The answer decides whether the next investment should go into better teaching content or into a friendlier welcome. The same question arises whenever information, training or communication is expected to change behaviour by changing what people know.

Mediation analysis addresses this by separating a total effect into an *indirect effect*, which passes through an intermediate variable called the mediator, and a *direct effect*, which covers every other route. Modern **causal mediation analysis** defines these quantities with potential outcomes, states the assumptions under which they can be estimated, and provides tools to check how sensitive the conclusions are to violations (Imai, Keele, & Tingley, 2010; VanderWeele, 2016). This article explains the approach, works through an illustrative example, and sets out practical guidance for survey and business research.

The classic approach and its limits

For decades, mediation was assessed with the "causal steps" of Baron and Kenny (1986). Using three regressions, the analyst checked that the treatment predicted the outcome, that the treatment predicted the mediator, and that the mediator predicted the outcome when the treatment was controlled for. If the treatment's coefficient shrank when the mediator was added, mediation was declared.

Two equations are central to this tradition. With treatment *T*, mediator *M* and outcome *Y*:

$$\mathrm{M} = \mathrm{i}_1 + \mathrm{aT} + \mathrm{e}_1$$

$$\mathrm{Y} = \mathrm{i}_2 + \mathrm{c}'\mathrm{T} + \mathrm{bM} + \mathrm{e}_2$$

The indirect effect is the product $a \times b$, and the direct effect is c'. In linear models without interaction, the total effect c equals $c' + ab$. Sobel (1982) derived a standard error for the product, but its sampling distribution is usually skewed, and bootstrap confidence intervals are now preferred (Preacher & Hayes, 2008).

The causal steps procedure has well-documented weaknesses. Simulation studies show that it has low power to detect mediation (MacKinnon et al., 2002). Requiring a significant total effect before testing for mediation is unnecessary, because direct and indirect effects can cancel out (Hayes, 2009; Zhao et al., 2010). Most importantly, the regressions describe associations. They do not show that the mediator *causes* changes in the outcome, even when the treatment itself is randomised (Bullock et al., 2010).

Defining direct and indirect effects causally

The causal approach starts from potential outcomes. Let $M(t)$ be the value the mediator would take under treatment t, and $Y(t, m)$ the outcome under treatment t and mediator value m. Robins and Greenland (1992) and Pearl (2009) defined two key quantities.

The **natural indirect effect** (NIE) is the change in the outcome when the treatment is held at its active level but the mediator moves from the value it would take without treatment to the value it takes with treatment:

$$\mathrm{NIE} = \mathrm{E}[\mathrm{Y}(1,\, \mathrm{M}(1)) - \mathrm{Y}(1,\, \mathrm{M}(0))]$$

The **natural direct effect** (NDE) is the change in the outcome when the treatment changes but the mediator is held at the value it would have taken without treatment:

$$\mathrm{NDE} = \mathrm{E}[\mathrm{Y}(1,\, \mathrm{M}(0)) - \mathrm{Y}(0,\, \mathrm{M}(0))]$$

These add up exactly to the total effect, TE = NDE + NIE, whatever the form of the model. Imai, Keele and Yamamoto (2010) call the indirect effect the *average causal mediation effect* (ACME). The definitions do not depend on linear regression, so the same quantities can be estimated for binary outcomes, counts or models with interactions between treatment and mediator (Valeri & VanderWeele, 2013). When there is no such interaction and the models are linear, the NIE equals the classic ab product.

The key assumption: sequential ignorability

Estimating natural effects requires stronger assumptions than estimating a total effect. Imai, Keele and Yamamoto (2010) express them as **sequential ignorability**:

1. The treatment is independent of the potential mediators and outcomes, given pre-treatment covariates. Randomisation of the treatment guarantees this.

2. The mediator is independent of the potential outcomes, given the treatment and pre-treatment covariates. In other words, there is no unmeasured confounding of the mediator–outcome relationship.

The second assumption is the crux. Randomising the treatment does *not* randomise the mediator. People who learn more from a tutorial may differ from those who learn less in prior skill, motivation or available time, and any of these may also affect the outcome. VanderWeele (2016) adds a further condition for natural effects: no confounder of the mediator–outcome relationship may itself be affected by the treatment. Conditioning on a mediator when such confounding exists can induce collider bias, so that even the direct effect is wrongly estimated (Cole & Hernán, 2002).

Because these assumptions cannot be tested with the observed data, a **sensitivity analysis** is an essential part of any causal mediation study (Imai, Keele, Tingley, & Yamamoto, 2011).

A worked example: does knowledge explain the effect of onboarding?

Consider a hypothetical software company that randomly assigns 2,000 new customers either to an interactive onboarding tutorial ($T = 1$) or to a standard welcome email ($T = 0$). Seven days later, all customers complete a short in-product quiz on the product's features, giving a knowledge score from 0 to 10 (M). The outcome is the number of days each customer is active in the first 90 days (Y). The model adjusts for pre-treatment covariates such as company size and plan type.

The illustrative estimates are:

- **Effect of the tutorial on knowledge:** $a = 1.6$ points on the quiz.
- **Effect of knowledge on activity, holding the tutorial constant:** $b = 2.5$ active days per quiz point.
- **Direct effect of the tutorial:** $c' = 2.0$ active days.

The indirect effect through knowledge is $1.6 \times 2.5 = 4.0$ active days, and the total effect is $2.0 + 4.0 = 6.0$ days. Two-thirds of the tutorial's effect on activity appears to work through what customers learn. A bootstrap 95% confidence interval for the indirect effect of, say, 3.1 to 4.9 days would indicate that the estimate is precise.

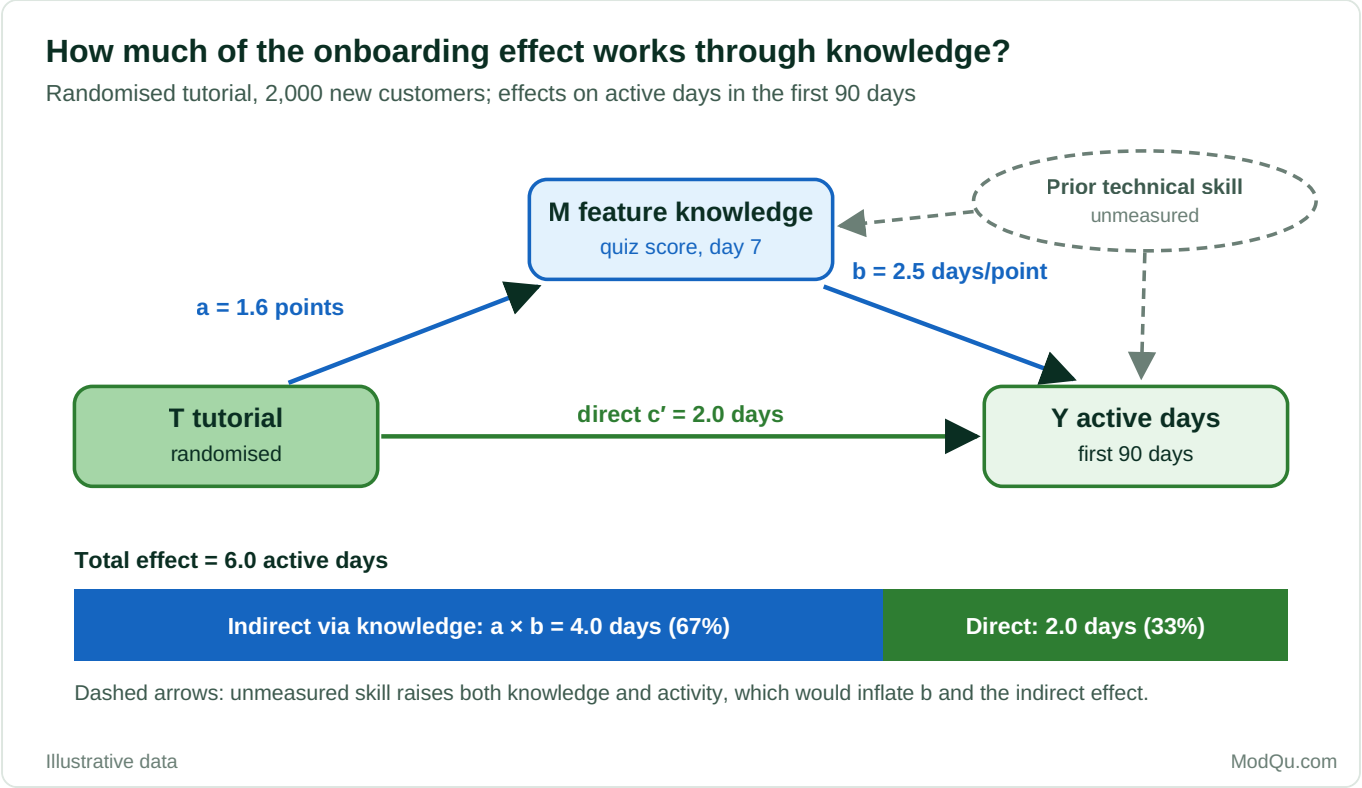


Figure 1. Decomposing the effect of an onboarding tutorial on 90-day activity into a pathway through feature knowledge and a direct pathway. The dashed node marks the mediator–outcome confounding that randomisation does not remove. Illustrative data; chart by ModQu.

The figure also shows why caution is needed. Customers with more prior technical skill are likely to score higher on the quiz *and* to use the product more, whatever their tutorial assignment. If skill is not measured, the coefficient *b* overstates the causal effect of knowledge, and the indirect effect is too large.

Sensitivity analysis

Imai, Keele and Yamamoto (2010) propose a sensitivity parameter ρ: the correlation between the error terms of the mediator and outcome equations. Under sequential ignorability, ρ = 0. A positive ρ means that unmeasured factors push the mediator and the outcome in the same direction, as prior skill would. The analyst recomputes the indirect effect over a range of ρ values and reports the value at which it would vanish. For linear models, the corrected indirect effect is

$$\mathrm{NIE}(\rho) = \mathrm{a} \times \left[\mathrm{b} - \rho\sigma_{\mathrm{Y}} \div \left(\sigma_{\mathrm{M}}\sqrt{(1-\rho^2)}\right)\right]$$

where σ_{M} and σ_{Y} are the residual standard deviations of the mediator and outcome models. In the example these are, illustratively, 2.0 quiz points and 15 active days.

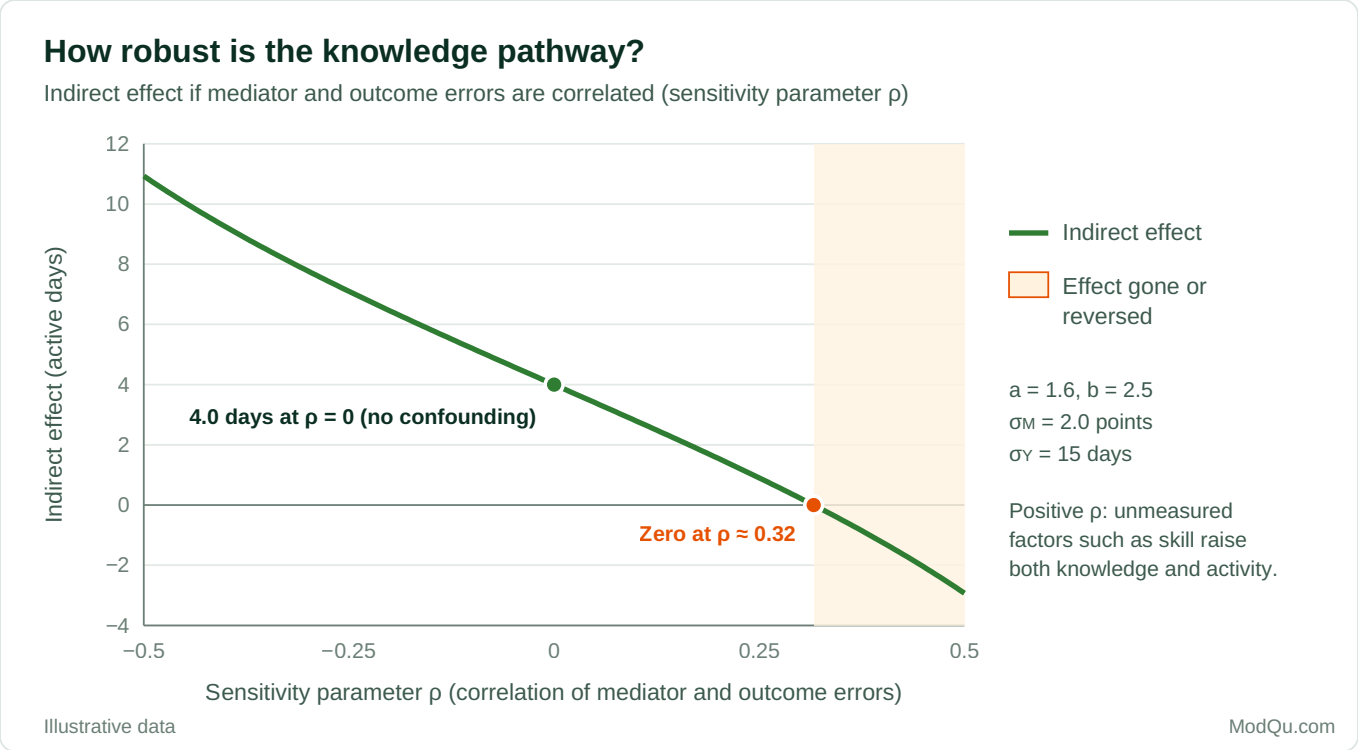


Figure 2. Sensitivity of the indirect effect to unmeasured mediator–outcome confounding. The estimate of 4.0 days at ρ = 0 falls to zero if the error correlation reaches about 0.32. Illustrative data; chart by ModQu.

In the illustration, the indirect effect disappears if the residual correlation reaches about 0.32. Whether that is plausible is a substantive judgement. If prior technical skill could easily create a correlation of that size, the knowledge pathway is fragile, and the company should measure skill before onboarding or test the mechanism directly, for example by randomising the content of the tutorial so that knowledge itself is manipulated.

Extensions

- **Treatment–mediator interaction.** The tutorial may make knowledge more valuable, for example because it shows users where to apply it. With interaction, the NDE and NIE are no longer simple coefficients, and the regression-based formulae of Valeri and VanderWeele (2013) or the simulation approach of Imai, Keele and Tingley (2010) should be used.
- **Binary and count outcomes.** For outcomes such as renewal or number of purchases, products of logistic or Poisson coefficients do not give the NIE. The potential outcomes definitions still apply, and software estimates them by simulation (Tingley et al., 2014).
- **Several mediators.** Knowledge may be one of several pathways, alongside satisfaction or habit formation. When mediators affect one another, separate path-specific effects require strong assumptions, and estimating the joint indirect effect through all mediators together is often more defensible (VanderWeele & Vansteelandt, 2014).
- **Timing.** Mediation is a process over time. Cross-sectional data, in which treatment, mediator and outcome are measured at the same moment, can give badly biased estimates of longitudinal mediation (Maxwell & Cole, 2007).

Practical guidance for survey and business data

- **Measure the mediator after the treatment and before the outcome.** In a survey, ask the knowledge questions after the exposure and measure the outcome later, or link to behavioural data collected afterwards.
- **Collect pre-treatment covariates that could confound the mediator and outcome.** Prior experience, baseline knowledge, motivation and demographics are typical candidates. These are the variables that make sequential ignorability credible.
- **Measure the mediator well.** Measurement error in a knowledge score biases the indirect effect, usually towards zero, and shifts part of it into the apparent direct effect. Use validated multi-item scales where possible.
- **Avoid post-treatment adjustment.** Do not control for variables that the treatment may have changed, other than the mediator itself.
- **Use survey weights and design-based inference** when the data come from a complex sample, and bootstrap in a way that respects the design.
- **Report the full decomposition.** Give the total, direct and indirect effects with confidence intervals, the proportion mediated with caution when the total effect is small, and the sensitivity analysis (VanderWeele, 2016).

Common mistakes

- **Treating the causal steps as proof of mechanism.** Significant paths show associations consistent with mediation, not a causal pathway (Bullock et al., 2010).
- **Assuming randomisation protects the mediator.** Only the treatment is randomised; mediator–outcome confounding remains.
- **Insisting on a significant total effect first.** Opposing pathways can hide a real indirect effect (Zhao et al., 2010).
- **Reporting a proportion mediated above 100% or below 0% without comment.** This happens when direct and indirect effects have opposite signs and should be explained rather than reported as a percentage.
- **Skipping sensitivity analysis.** Without it, readers cannot judge how robust the decomposition is.

Key takeaways

Causal mediation analysis splits an effect into the part that runs through a mediator, such as knowledge, and the part that does not. The natural direct and indirect effects are defined with potential outcomes and add up to the total effect for any model. Their estimation rests on sequential ignorability, and above all on the absence of unmeasured mediator–outcome confounding, which randomising the treatment does not guarantee. Good studies measure the mediator at the right time, collect the covariates that could confound it, use bootstrap or simulation-based intervals, and report a sensitivity analysis alongside the estimates (MacKinnon et al., 2007). Used in this way, mediation analysis turns the question "does it work?" into the more useful question "how does it work?".

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98*(4), 550–558. https://doi.org/10.1037/a0018933

Cole, S. R., & Hernán, M. A. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology, 31*(1), 163–165. https://doi.org/10.1093/ije/31.1.163

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408–420. https://doi.org/10.1080/03637750903310360

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309–334. https://doi.org/10.1037/a0020761

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105*(4), 765–789. https://doi.org/10.1017/S0003055411000414

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science, 25*(1), 51–71. https://doi.org/10.1214/10-STS321

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*(1), 593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*(1), 83–104. https://doi.org/10.1037/1082-989X.7.1.83

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods, 12*(1), 23–44. https://doi.org/10.1037/1082-989X.12.1.23

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology, 3*(2), 143–155. https://doi.org/10.1097/00001648-199203000-00013

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology, 13*, 290–312. https://doi.org/10.2307/270723

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software, 59*(5), 1–38. https://doi.org/10.18637/jss.v059.i05

Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods, 18*(2), 137–150. https://doi.org/10.1037/a0031034

VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health, 37*(1), 17–32. https://doi.org/10.1146/annurev-publhealth-032315-021402

VanderWeele, T., & Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods, 2*(1), 95–115. https://doi.org/10.1515/em-2012-0010

Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research, 37*(2), 197–206. https://doi.org/10.1086/651257