Clarifying Causal Mediation Analysis for the Applied Researcher: Defining Effects Based on What We Want to Learn

Presented by Yuqi Li

Duke B&B

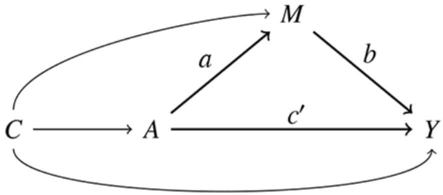
February 20, 2026

Nguyen, Schmid & Stuart (2021) — Psychological Methods, 26(2): 255–271

Basic Setup and Notation

Variables:

- A = exposure / treatment (binary here: $0, 1$)
- M = mediator
- Y = outcome
- C = baseline covariates (pre-exposure)



Two causal pathways:

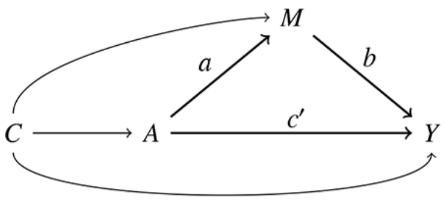
- Direct path: $A \to Y$
- Mediated path: $A \to M \to Y$

Mediation: What It Asks and Why It Matters

Core question: When an exposure A changes an outcome Y, *how much of that change operates through* an intermediate variable M?

Basic story:

- A may affect Y *directly*
- A may affect Y *indirectly* by changing M



Why it matters (research significance):

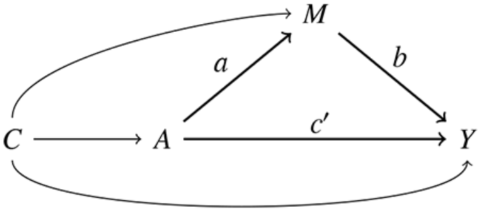
- Mechanism: helps explain *why* an intervention works
- Design: points to modifiable targets (M) for improving outcomes
- Policy: separates what would change under different intervention components

Traditional Product Method

Linear model (classical setup)

$$M \leftarrow A + C$$

$$Y \leftarrow A + M + C$$



Product of coefficients idea

- Indirect $\approx (A \to M) \times (M \to Y) = ab$
- Direct $\approx$ coefficient of A in the Y model $= c'$

Assumes linearity and no $A \times M$ interaction

Why this can fail:

- Nonlinear models: coefficients are not on an additive causal scale
- $A \times M$ interaction changes the decomposition
- Model misspecification changes the target quantity

Key message: Traditional mediation defines the effect through a model. Causal mediation defines the causal effect first, then chooses a model to estimate it.

Three-Step Causal Workflow

1. **Define the estimand (model-free)**
   Specify a causal contrast using potential outcomes (what world vs what world?). Decide whether you want *explanatory* (natural effects) or *policy* (interventional effects).

2. **Identify (assumptions $\rightarrow$ observable functional)**
   State the causal assumptions (confounding control, temporality, no intermediate confounding if needed). Determine whether the estimand can be written in terms of the observed data distribution.

3. **Estimate (choose statistical tool)**
   Pick an estimator consistent with identification (g-formula, weighting, doubly robust). Report uncertainty and do sensitivity analysis for unmeasured confounding.

Definition $\rightarrow$ Identification $\rightarrow$ Estimation

Two Perspectives: Explanatory vs. Interventional

Explanatory (mechanism)

- Goal: explain the total effect
- Natural effects: NDE/NIE
- Requires cross-world assumptions

Interventional (policy)

- Goal: "What if we change M?"
- Interventional effects (later)
- Often identified with weaker requirements

Different questions $\rightarrow$ different estimands $\rightarrow$ different assumptions

Total Effect (TE): Individual vs. Population

Individual total effect:

$$TE_i := Y_i(1) - Y_i(0)$$

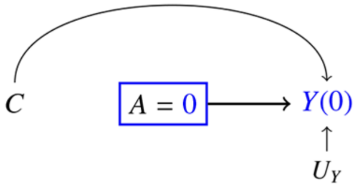
Population (average) total effect:

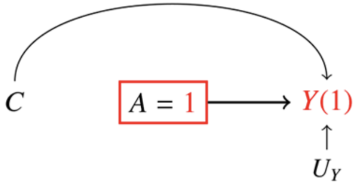
$$TE := E[Y(1)] - E[Y(0)]$$

Interpretation: change in outcome if everyone were exposed vs. if everyone were unexposed

Baseline identification idea (given C):

- No unmeasured A–Y confounding given C
- Positivity and consistency





Potential Outcomes: Where Natural Effects Come From

Counterfactual variables:

- $Y(a)$ = outcome if $A = a$
- $M(a)$ = mediator if $A = a$
- $Y(a, m)$ = outcome if $A = a$ *and* $M = m$

Natural effects use a cross-world term:

$$Y(1, M(0)) \quad \text{(treated exposure, control-level mediator)}$$

Why this is hard:

- For any individual, we never observe both $M(0)$ and $M(1)$
- Identification needs stronger, untestable assumptions than total effects

Natural Effects: Decomposing the Total Effect

Start from the total effect contrast:

$$TE = E[Y(1, M(1))] - E[Y(0, M(0))]$$

Insert an "in-between" world by adding and subtracting $E[Y(1, M(0))]$**:**

$$TE = \underbrace{E[Y(1, M(1))] - E[Y(1, M(0))]}_{\text{NIE}(1\cdot)} + \underbrace{E[Y(1, M(0))] - E[Y(0, M(0))]}_{\text{NDE}(\cdot 0)}$$

Information-flow metaphor: Switching $A : 0 \to 1$ sends "information" along two paths. *Freeze the mediator path first* (keep M at $M(0)$) to isolate the direct-path change.

Natural Direct and Indirect Effects (NDE & NIE)

Natural direct effect (NDE):

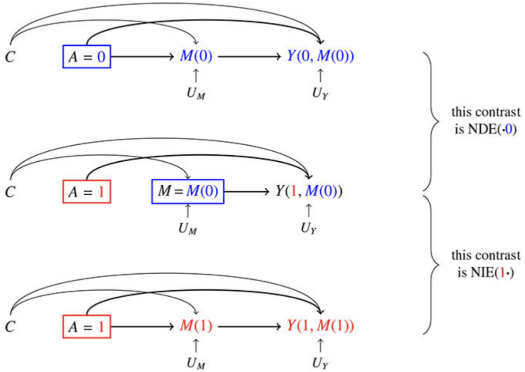
$$NDE(\cdot 0) := E[Y(1, M(0))] - E[Y(0, M(0))]$$

Interpretation: change A, but keep M at its natural control level.

Natural indirect effect (NIE):

$$NIE(1\cdot) := E[Y(1, M(1))] - E[Y(1, M(0))]$$

Interpretation: keep $A = 1$, let M switch from $M(0)$ to $M(1)$.



Important note: with $A \times M$ interaction, there are *two* NDEs and *two* NIEs (order of decomposition matters).

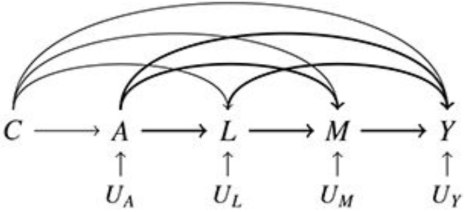
The L Problem: Intermediate Confounding

Setup:
Suppose a variable L:

- is affected by exposure A
- affects both mediator M and outcome Y

Example:

- Intervention $\rightarrow$ stress (L)
- Stress $\rightarrow$ parenting (M)
- Stress and parenting $\rightarrow$ child outcome (Y)



Why Natural Effects Are Not Identified with L

Natural effects require quantities like:

$$Y(1, M(0))$$

But with intermediate confounder L:

- L depends on A
- L affects both M and Y
- We would need $L(1)$ and $L(0)$ simultaneously

Problem:

- Cross-world counterfactuals
- Not identified without strong assumptions

Natural direct/indirect effects generally fail when L exists.

Core Idea: Replace Individual $M(0)$ with a Distribution

Instead of fixing mediator to each person's unobserved $M(0)$,

We intervene on the **mediator distribution**:

$$\mathcal{M}(0 \mid C) \sim p(M \mid A = 0, C)$$

Meaning:

- For individuals with covariates C
- Randomly assign mediator values drawn from control group distribution

This avoids cross-world individual-specific counterfactuals.

Interventional Direct and Indirect Effects

Interventional Direct Effect (IDE):

$$\mathrm{IDE}(\cdot 0) = E[Y(1, \mathcal{M}(0|C))] - E[Y(0, \mathcal{M}(0|C))]$$

Effect of shifting A from 0 to 1
while keeping mediator distribution fixed at control.

Interventional Indirect Effect (IIE):

$$\mathrm{IIE}(0 \cdot) = E[Y(0, \mathcal{M}(1|C))] - E[Y(0, \mathcal{M}(0|C))]$$

Effect of shifting mediator distribution
while keeping $A = 0$ fixed.

Natural vs Interventional Effects

- Natural effects use $M(0)$ for each individual
- Interventional effects use distribution $p(M|A = 0, C)$

Consequences:

- Natural effects decompose TE
- Interventional effects generally do NOT
- Natural effects require no intermediate confounder
- Interventional effects allow L

Interventional effects answer "what if we changed the mediator distribution?"

Example: Sexual Minority Disparities

Population: Adolescents

- $A = 1$: sexual minority
- $A = 0$: sexual majority

Mediator:

- M = bullying experience

Outcome:

- Y = well-being

Covariates:

- C = demographics not affected by A

Observed disparity:

$$\text{disparity}(C) = E[Y|A=1, C] - E[Y|A=0, C]$$

What If We Equalized Bullying?

Question: How much of the well-being disparity would disappear if sexual minority youth had the same bullying distribution as sexual majority youth?

Intervention:
Replace minority bullying distribution with $d_{M0,C} = p(M|A=0,C)$

Decomposition:

$$\text{disparity}(C) = \underbrace{E[Y|A=1,C] - E[Y(1,\mathcal{M}_{0,C})|A=1,C]}_{\text{disparity removed}} + \underbrace{E[Y(1,\mathcal{M}_{0,C})|A=1,C] - E[Y|A=0,C]}_{\text{remaining disparity}}$$

Key insight:
First term is a causal interventional effect. Second term is not causal (cross-group contrast).

Controlled Direct Effect (CDE)

Idea: Fix the mediator to a constant level m for everyone.

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

Interpretation: Effect of exposure when the mediator is held at m.

When meaningful?

- When a structural intervention can set $M = m$ for all.
- Example: law fixing water heater temperature to 120°F.

CDE is an interventional effect.

Example: Water Heater Regulation

Child injury prevention program works by lowering water temperature.

Original mechanism: Program $\rightarrow$ parental awareness $\rightarrow$ lower temperature $\rightarrow$ fewer burns

City policy: Law sets maximum water temperature to 120°F.

Question: Would the program still reduce burns if temperature is already fixed?

This corresponds to: CDE($120°F$)

If instead the policy only reduces temperature on average (not fixed exactly), we move from CDE to a GIDE with a temperature distribution D.

Generalized Interventional Direct Effect (GIDE)

We are not restricted to fixing M to a single value.

Let D be **any** mediator distribution.

$$\text{GIDE}(\cdot D) = E[Y(1, \mathcal{M}_D)] - E[Y(0, \mathcal{M}_D)]$$

where $\mathcal{M}_D$ is a random draw from distribution D.

Examples of D:

- Control distribution $p(M|A = 0, C)$
- Treated distribution $p(M|A = 1, C)$
- 50-50 mixture
- Any policy-target distribution

Key insight: CDE is a special case of GIDE where D is a single point mass at m.

Identification Depends on the Target Estimand

Core principle: Identification follows from the intervention being contrasted.

Estimand	**No A–Y conf.**	**No A–M conf.**	**No M–Y conf.**	**No L**	**Cross-world**
TE	✓				
CDE	✓		✓		
NDE/NIE	✓		✓	✓	✓
IDE/IIE	✓	✓	✓		

Important clarifications:

- Different causal questions imply different identifying assumptions.
- Cross-world assumptions appear only for natural effects.

Estimation: What Must Be Modeled?

After defining the estimand and assumptions, estimation becomes mechanical.

Three core components in mediation settings:

1. Outcome model: $E[Y \mid A, M, C]$
2. Mediator model: $p(M \mid A, C)$
3. Exposure model: $p(A \mid C)$

Main estimation strategies:

- G-computation (model-based plug-in)
- Inverse Probability Weighting (reweighting)
- Doubly robust methods (combine two models)

With intermediate confounding L:

- Additional modeling is required
- Natural effects may not be identified
- Interventional effects often remain feasible

Sensitivity Analysis: The Inevitable Step

Key reality: The mediator is rarely randomized.

- Unmeasured M–Y confounding is the main vulnerability
- Identification assumptions are not testable from data
- Results depend on structural assumptions

Good practice:

- Report sensitivity analysis for mediator–outcome confounding
- Be explicit when using IDE/IIE as approximations to NDE/NIE
- Separate causal claims from descriptive contrasts

Beyond the Basic Mediation Setting

Active research areas include:

- Survival outcomes and time-to-event mediation
- Multiple mediators
- Time-varying exposures and mediators
- Dynamic treatment regimes
- Policy-based distributional interventions

Big-picture message:

Choose the estimand to match the scientific question.
Identification and estimation follow from that choice.

Practical Workflow (and What to Remember)

1. **Start from the causal question (not association).**
   Decide: *Explanatory* (decompose TE) vs. *Interventional* (policy on M).

2. **Write the estimand that matches the question.**
   Use potential outcomes notation and name the target effect (TE, NDE/NIE, IDE/IIE, CDE, GIDE).

3. **Use the DAG to state identification assumptions.**
   Be explicit about confounding and whether intermediate confounders L are present (natural effects add cross-world requirements).

4. **Estimate, report uncertainty, and check robustness.**
   Fit an estimator consistent with the assumptions and include sensitivity analysis.

Core message: Your question $\rightarrow$ your estimand. Everything else follows.