Testing Full Mediation of Treatment Effects and the Identifiability of Causal Mechanisms

Martin Huber Kevin Kloiber Lukáš Lafférs

University of Fribourg · University of Munich · Matej Bel University / NHH

NBS Bratislava 2026

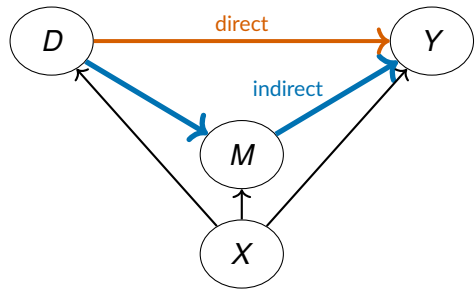
What is mediation analysis?

A treatment D affects an outcome Y…

…but **how**? Through what channel?

> **Mediation analysis** decomposes the total effect of D on Y into:
>
> ▶ indirect effect - operating through M
>
> ▶ direct effect - bypassing M

Full mediation: the entire effect flows through M — no direct path $D \to Y$



D = treatment M = mediator

Y = outcome X = covariates

Why does it matter? — Examples

Treatment *D*	**Mediator** *M*	**Outcome** *Y*
Job-training programme	Employment / skills acquired	Long-run wages
Early childhood intervention	Cognitive skills at age 5	Adult earnings
Minimum wage increase	Firm automation	Low-skill employment
Remote work policy	Work-life balance	Worker productivity
Union membership	Bargaining power	Wage premium
Immigration inflow	Labor supply competition	Native wages
Childcare subsidy	Reduction in unpaid care work	Female LF participation
Education reform	School attainment / skills	Lifetime earnings
CBT for maternal depression	Mental health, family support	Child development
Peers' attitude info (Saudi Arabia)	Job-search service sign-up	Wife's employment

Why mechanisms matter

- scale-up: replicate by targeting M
- policy design: intervene on the right thing
- theory testing: confirm hypothesised channel

The identification challenge

- M is rarely randomised
- unobserved M–Y confounders bias estimates
- hard to test even with randomised D

Mediation questions you know

Setting	Treatment → mediator → outcome	Familiar tool
ALMP	training → re-employment → earnings	LATE, Heckman selection
Returns to schooling	education → occupation/skills → wages	Card, Mincer decomp.
Wage gap decomp.	group → characteristics → wage	Oaxaca–Blinder
Monetary transmission	policy rate → bank lending / expectations → output	VAR, narrative IV
RCT analysis	treatment arm → behavior → outcome	IV, control function

Most existing tools either **decompose under strong functional form** (Oaxaca–Blinder) or **require an instrument for** M (rare).
We ask: *is the full-mediation story even consistent with the data?*

Running example — job training

Setup (JTPA-style)

D	randomised offer of training
M	re-employment within 12 months
Y	cumulative earnings, 4 years out
X	age, education, prior earnings, …

Total effect of D on Y is identified by randomisation.

But *how does training raise earnings?*

Two competing stories

- **Placement channel only:** training pushes people into jobs sooner; once employed, earnings track the same path
- **Direct human-capital effect:** training also raises productivity / wages *conditional on* working

> **Policy stake.** If the channel is purely placement, a cheap job-search service may substitute for (expensive) training. If there is a direct skill effect, it cannot.

A joint test of

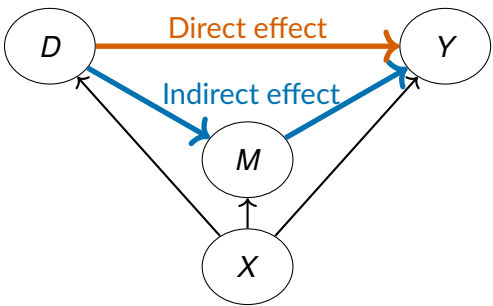
full mediation

and

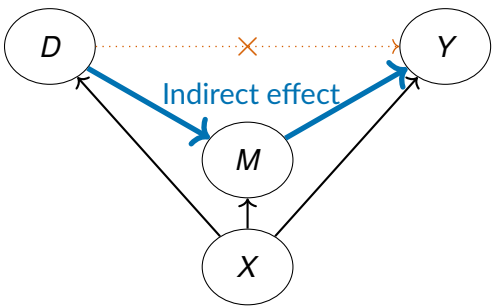
identifiability of causal mechanisms

in mediation analysis

Decomposition of effects



Full mediation condition

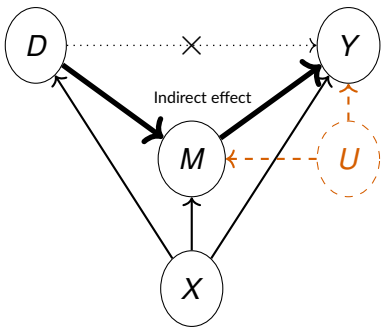


Why testing full mediation is hard

Naive approach: regress Y on D and M, test $\hat{\beta}_D \approx 0$

Problem: if M is endogenous, conditioning on M opens a collider path — spurious D–Y association
($D \rightarrow M \leftarrow U \rightarrow Y$)

Bias arises *even when D is randomized*



The bad-control problem in our language

Linear DGP with full mediation *but* unobserved M–Y confounder U:

$$M = \alpha D + U + \varepsilon_M$$
$$Y = \beta M + \delta U + \varepsilon_Y \qquad (\text{no } D \text{ in } Y)$$

Run the **naive** Baron–Kenny regression Y on D, M:

$$\hat{\beta}_D \xrightarrow{p} -\frac{\alpha \delta \sigma_U^2}{\sigma_M^2} \neq 0$$

Even with randomised D, conditioning on the mediator opens the collider $D \to M \leftarrow U \to Y$.

Naive test concludes: *direct effect exists.*
Truth: full mediation holds; M is endogenous.

Angrist–Pischke would say…

- M is a **bad control**: post-treatment, endogenous
- Sign and size of bias depend on δ, σ_U^2 — not on the true mediation structure
- No reason to think $\hat{\beta}_D \approx 0$ implies full mediation

Our test: a rejection comes from a direct effect *or* from M-endogeneity. A non-rejection rules out both.

Literature

- mediation analysis
  Robins and Greenland (1992), Robins (2003), Pearl (2001), Tchetgen Tchetgen and Shpitser (2012)
- problems
  Acharya, Blackwell, and Sen (2016)
- testing
  Huber and Kueck (2022)
- full mediation
  Kwon and Roth (2026) More

Contribution

- joint test for full mediation and mediator exogeneity
- clarify the relationship between assumptions
- DML implementation, simulations, applications

Identification

Setup

Observed variables

D	treatment $\{0,1\}$
M	mediator
Y	outcome
X	covariates

Potential outcomes

$M(d)$	mediator under d
$Y(d,m)$	outcome under (d,m)
$Y(m)$	under full mediation

Assumptions

1	causal structure	excl. restrictions
2	common support	$p(D,M\|X) \in (0,1)$
3	first stage	$\Pr(M(1) \neq M(0)) > 0$
4	random treatment	$D \perp\!\!\!\perp Y(d,m), M(d) \mid X$
4a	exog. treatment (wrt Y)	$D \perp\!\!\!\perp Y(d,m) \mid X$
4b	exog. treatment (wrt M)	$D \perp\!\!\!\perp M(d) \mid X$
5	cond. full mediation	$Y(d,m) = Y(m) \mid X$ a.s.
6	exog. mediator	$M \perp\!\!\!\perp Y(m) \mid X$

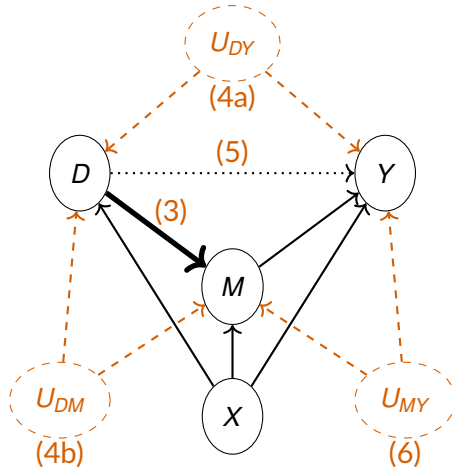
Identifying Assumptions

Assumption 1

- $M(y) = M$:
  ($Y \not\to M$, Y does not cause M)
- $D(m, y) = D$
- $X(d, m, y) = X$
- Faithfulness

Assumption 3

- $D \to M$ exists



The key testable implication (TI)

$$Y \perp\!\!\!\perp D \mid M, X$$

Outcome independent of treatment, conditional on mediator and covariates

What is the test really doing?

Once we know the mediator M and covariates X,
does the treatment D still help predict the outcome Y?

Reject (TI)

D still predicts Y given (M, X). **Either:**

- a direct path $D \to Y$ exists, **or**
- M is endogenous (collider bias), **or both**

We cannot separate the two — but the full-mediation story is incomplete either way.

Do not reject (TI)

Consistent with full mediation **and** mediator exogeneity holding jointly.

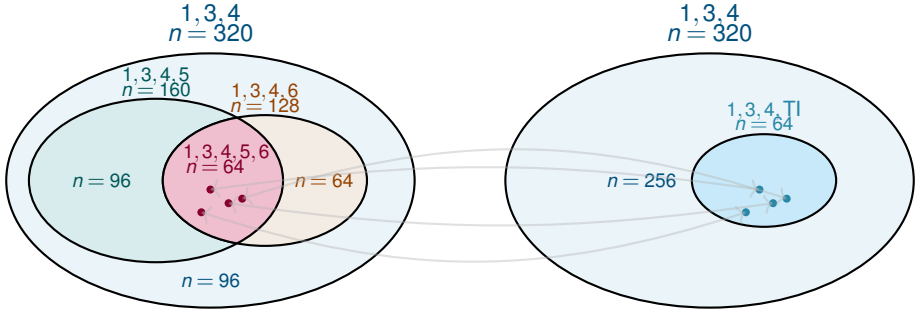
- does *not prove* either holds (power)
- but it is the data signature you would expect

Theorem 1 — randomly assigned treatment

Under Assumptions 1, 3, 4 (CI of treatment):

$$\underbrace{5}_{\substack{\text{cond. full}\\ \text{mediation}}}\ ,\ \underbrace{6}_{\substack{\text{CI of}\\ \text{mediator}}} \iff \underbrace{(\text{TI})}_{\substack{\text{testable}\\ \text{implication}}}$$

- One CI test jointly verifies **full mediation** and **mediator exogeneity**
- Rejection $\Rightarrow$ at least one assumption fails



Under Assumptions 1, 3, 4 :

$$\underbrace{5}_{\text{cond. full mediation}}\ ,\ \underbrace{6}_{\text{CI of mediator}} \iff \underbrace{5,6}_{\substack{\text{cond. full mediation} \\ \text{CI of mediator}}} \iff \underbrace{(\text{TI})}_{\substack{\text{testable} \\ \text{implication}}}$$

▸ Visual (appendix)

Theorem 2 — observational data

Under Assumptions 1, 3 only — no treatment randomization assumed:

$$\underbrace{4a}_{\substack{\text{no direct}\\ \text{D–Y conf.}}}\ ,\ \underbrace{5}_{\substack{\text{cond. full}\\ \text{mediation}}}\ ,\ \underbrace{6}_{\substack{\text{CI of}\\ \text{mediator}}} \iff \underbrace{(\text{TI})}_{\substack{\text{testable}\\ \text{implication}}}$$

- Full mediation and M-exogeneity remain testable
- Treatment–mediator (4b) confounding is **not** detected by (TI)
- Non-rejection $\not\Rightarrow$ indirect effect is identified

Relation to front-door / back-door

$$\underbrace{\Pr(Y{=}y \mid D{=}d, X)}_{\text{back-door}} = \underbrace{\sum_m \Pr(M{=}m \mid D{=}d, X) \sum_{d'} \Pr(Y{=}y \mid D{=}d', M{=}m, X)\, \Pr(D{=}d' \mid X)}_{\text{front-door}}$$

(BD=FD)

Theorem 3. $(\text{TI}) \Rightarrow (\text{BD} = \text{FD})$

Theorem 4. $(\text{BD} = \text{FD})$ + Separability $\Rightarrow$ (TI)

BD–FD equivalence is a *weaker* test than (TI)

Separability: $\Pr(Y{=}y \mid D, M, X) = \alpha(D, X) + \beta(M, X)$
rules out D–M interactions; strong assumption

How this fits with others

Approach	Key requirement	What it delivers
IV for M	valid IV $Z \to M$, excl. from Y	point ID of indirect effect
Control function	structural model of M, dist. assumption on errors	point ID under restrictions
Sensitivity (Imai–Keele)	bound on M–Y confounding	bounds on indirect effect
Front-door (Pearl)	full mediation *assumed*	point ID of total / indirect effect
Partial ID (Kwon & Roth)	monotonicity, finite M	sharp bounds; sharp test of full med
Our test (TI)	randomised D + Asm. 1, 3 (weaker without random. — Thm. 2)	joint test of full med. & M-exog.; DML, high-dim X

Implementation

Double Machine Learning

$H_0: \; E[Y \mid M, X, D{=}1] - E[Y \mid M, X, D{=}0] = 0$

Doubly robust score (Apfel et al. 2023; Chernozhukov et al. 2018)

$$\tilde{\psi} = (\mu_1 - \mu_0)^2 + 2(\mu_1 - \mu_0)\left[\frac{(Y-\mu_1)D}{p} - \frac{(Y-\mu_0)(1-D)}{1-p}\right] + (\mu_1 - \mu_0) + \left[\frac{(Y-\mu_1)D}{p} - \frac{(Y-\mu_0)(1-D)}{1-p}\right] - \theta$$

$\mu_d = E[Y \mid M, X, D{=}d], \quad p = \Pr(D{=}1 \mid M, X)$

- **Neyman orthogonality** — insensitive to nuisance estimation error
- **5-fold cross-fitting** — avoids overfitting bias
- **Lasso / ML nuisances** — $\sqrt{n}$-consistent, asymptotically normal

Why mean independence?

$H_0 : \; E[Y \mid M, X, D{=}1] - E[Y \mid M, X, D{=}0] = 0$

Why mean independence instead of full independence?

It is practical:

- **ATE** — for average treatment effects, mean independence suffices
- **full CI** — requires non-parametric distributional tests, more demanding
- **mean independence** — enables the doubly robust DML framework

Simulation

Simulation design

$$D = \mathbf{1}\{X'\beta + U_1 + \lambda U_2 > 0\} \quad \leftarrow D\text{–}M \text{ confounding if } \lambda > 0$$
$$M = 0.5\,D + X'\beta + \delta\, U_1 + U_2$$
$$Y = M + X'\beta + \gamma D + \delta\, U_1 + U_3$$

	Parameters	Violation tested
Null holds	$\delta = 0,\ \gamma = 0$	—
Null fails	$\delta \neq 0$	D–M–Y confounding
Null fails	$\gamma \neq 0$	direct effect $D \to Y$

$p = 200$ covariates · $n \in \{1000, 4000\}$ · 1000 replications · lasso, 5-fold cross-fitting
Theorem 1: $\lambda = 0$ · Theorem 2: $\lambda = 0.25$ (introduces D–M confounding)

Results — rejection rates at the 5% level

	Scenario	$n = 1{,}000$	$n = 4{,}000$
Main design, $\lambda = 0$			
Null	$\delta = 0,\ \gamma = 0$	0.055	0.040
M–Y conf.	$\delta = 0.25,\ \gamma = 0$	0.969	**1.000**
Direct effect	$\delta = 0,\ \gamma = 0.2$	0.532	**1.000**
With D–M confounding, $\lambda = 0.25$			
Null	$\delta = 0,\ \gamma = 0$	0.030	0.037
M–Y conf.	$\delta = 1,\ \gamma = 0$	**1.000**	**1.000**
Direct effect	$\delta = 0,\ \gamma = 0.2$	0.482	**0.993**

D–M confounding does not inflate size (Theorem 2 confirmed) · Strong power at $n = 4{,}000$

Empirical Illustrations

Social norms — Bursztyn et al. (2020), Saudi Arabia

- men underestimate how open are other men about women working outside of the home
- experiment: receive info on other men's beliefs or not
- job-search service or gift card
- treated men's women prob of applying for job-interview higher within 6 months
- is this short-run sign-up for job-search service the main driver of the outcomes in the longer-term?
- $n = 284$

Bursztyn, L., Gonzalez, A. L., and Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *American Economic Review*, 110(10), 2997–3029.

- D Information on peers' attitudes
- M Job-matching service sign-up
- Y Wife applied for outside job

Controls: baseline beliefs, employment, education, demographics

Mediator	p-value
Job-matching	0.004 **
Kwon & Roth (2026): $p = 0.020$	

Stronger rejection may also reflect mediator endogeneity

Perinatal depression — Baranov et al. (2020), Pakistan

- RCT of CBT to reduce depression for pregnant women in rural Pakistan
- studied effect on financial empowerement
- grandmother presence (proxy for family support) and relationship with husband as mediators
- $n \sim 600$

Baranov, V., Bhalotra, S., Biroli, P., and Maselko, J. (2020). Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial. *American Economic Review*, 110(3), 824–859.

- D CBT intervention (Thinking Healthy)
- M Grandmother presence; husband relationship quality
- Y Financial empowerment (7 yrs later)

Controls: age, education, employment, depression severity, wealth, family structure,...

Mediator	p-value
Grandmother	0.011 *
Relationship	0.022 *
Both	0.013 *

Kwon & Roth (2026): jointly $p = 0.654$

A mechanism is missing, or mediators are endogenous

Before running the test

- Is D randomised or quasi-experimental?
  Thm. 1 (random D) or Thm. 2 (observational)
- Is M measured *after* D but *before* Y?
- Are there strong priors on a direct $D \to Y$ path or M–Y confounders?
- Check overlap of (D, M, X)

Reading the result

- **Reject:** mechanism story is incomplete — direct effect *or* M endogeneity (or both)
- **Don't reject:** data consistent with full mediation *and* M-exogeneity; proceed with decomposition
- **Large policy stakes:** combine with sensitivity analysis

Before claiming *"the effect works through M"*:

check that your data is consistent with that story.

Conclusion

Summary

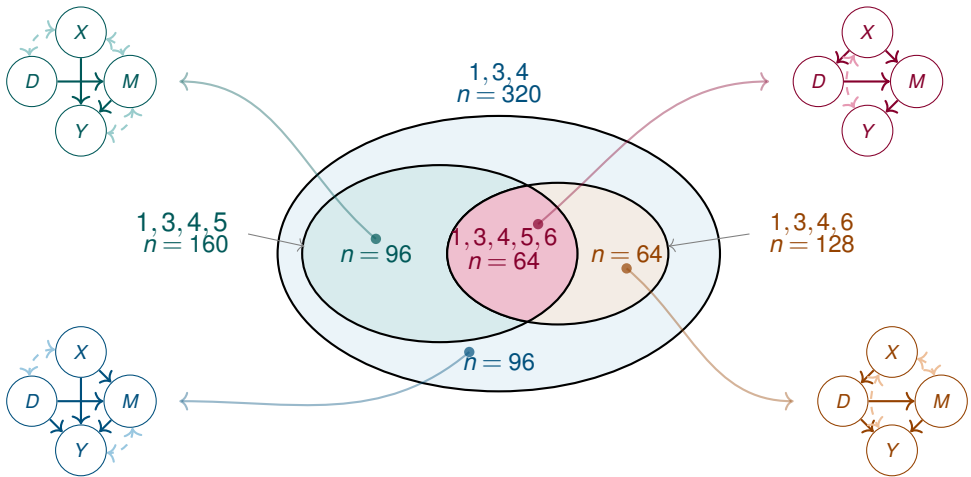
$Y \perp\!\!\!\perp D \mid M, X$ jointly tests **full mediation** and **mediator exogeneity**

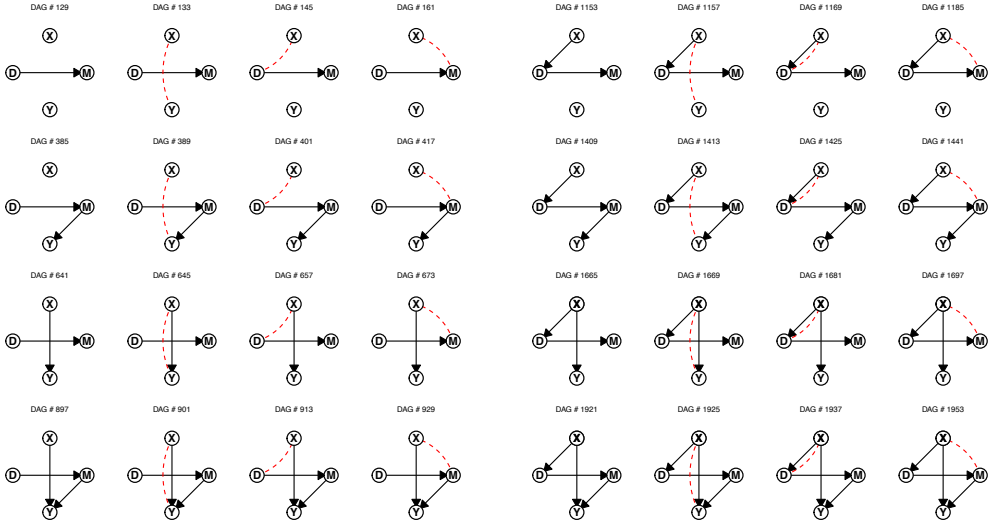
- **Randomized D:** full mediation *and* identifiability of indirect effects jointly testable
- **Observational D:** full mediation still testable; D–M confounding is not
- **BD = FD** is weaker — implies (TI) only under separability
- **DML:** $\sqrt{n}$-consistent, doubly robust, high-dimensional X
- Both applications **reject** the joint null

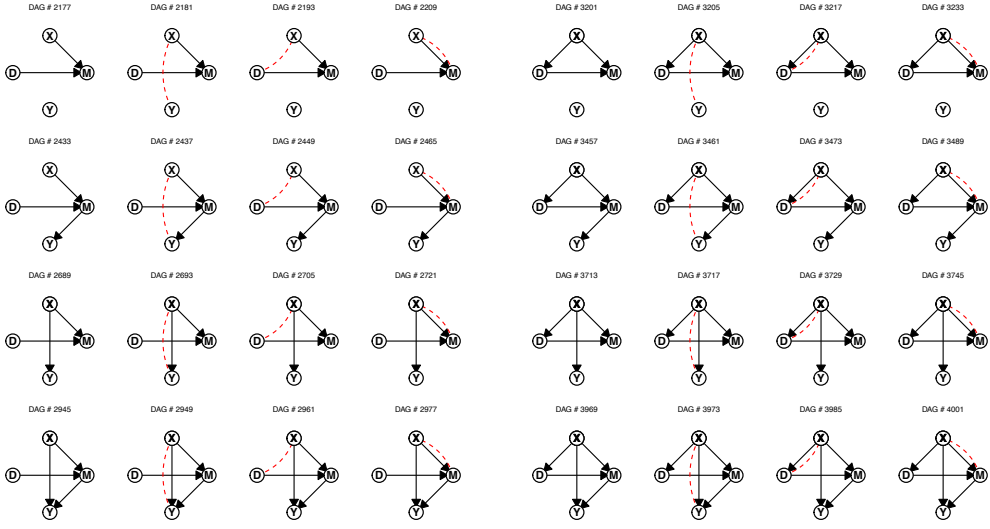
References

- Acharya, A., Blackwell, M., and Sen, M. (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, 110, 512–529.
- Apfel, N., Hatamyar, J., Huber, M., and Kueck, J. (2023). Learning control variables and instruments for causal analysis in observational data. Working paper, University of Fribourg.
- Baranov, V., Bhalotra, S., Biroli, P., and Maselko, J. (2020). Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial. *American Economic Review*, 110(3), 824–859.
- Bursztyn, L., Gonzalez, A. L., and Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *American Economic Review*, 110(10), 2997–3029.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.
- Huber, M. and Kueck, J. (2022). Testing the identification of causal effects in observational data. arXiv preprint arXiv:2203.15890.
- Kwon, S. and Roth, J. (2026). Testing mechanisms. *The Review of Economic Studies*
- Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420. Morgan Kaufmann, San Francisco.
- Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. J. Green, N. L. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, 70–81. Oxford University Press.
- Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3, 143–155.
- Tchetgen, E. J. T., & Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3), 1816.

Theorem 1 — visual (all DAG classes)




DAG # 129
DAG # 133
DAG # 145
DAG # 161
DAG # 1153
DAG # 1157
DAG # 1169
DAG # 1185
DAG # 385
DAG # 389
DAG # 401
DAG # 417
DAG # 1409
DAG # 1413
DAG # 1425
DAG # 1441
DAG # 641
DAG # 645
DAG # 657
DAG # 673
DAG # 1665
DAG # 1669
DAG # 1681
DAG # 1697
DAG # 897
DAG # 901
DAG # 913
DAG # 929
DAG # 1921
DAG # 1925
DAG # 1937
DAG # 1953
X
D
M
Y


DAG # 2177
DAG # 2181
DAG # 2193
DAG # 2209
DAG # 3201
DAG # 3205
DAG # 3217
DAG # 3233
DAG # 2433
DAG # 2437
DAG # 2449
DAG # 2465
DAG # 3457
DAG # 3461
DAG # 3473
DAG # 3489
DAG # 2689
DAG # 2693
DAG # 2705
DAG # 2721
DAG # 3713
DAG # 3717
DAG # 3729
DAG # 3745
DAG # 2945
DAG # 2949
DAG # 2961
DAG # 2977
DAG # 3969
DAG # 3973
DAG # 3985
DAG # 4001
X
D
M
Y

- 34752 DAGs with obs. variables (X, D, M, Y) and pairwise confounders
- 4096 that satisfy Assumption 1
- 2048 that satisfy Assumption 1 and 3
- 320 that satisfy Assumption 1, 3 and 4
- 64 that satisfy Assumption 1, 3, 4, 5 and 6

Back to Theorem 1

Comparison to Kwon and Roth (2026)

Kwon and Roth (2026):

- tests sharp null of full mediation
- partial identification approach
- sharp bounds via linear program
- discretization of cont. variables, M has finite support
- monotonicity assumptions
- degree of full mediation violations
- provide several modifications (e.g. violations of monotonicity)
- many covariates X

Kwon, S. and Roth, J. (2026). Testing mechanisms. *The Review of Economic Studies*

Back to Literature