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

Martin Huber   Kevin Kloiber   Lukáš Lafférs

Uni Fribourg · Uni Copenhagen · IER SAV / NHH

SEAM · NOeG · Vienna 2026

A joint test of

full mediation

and

identifiability of causal mechanisms

in mediation analysis

Mediation questions you know

Setting	Treatment $\rightarrow$ mediator $\rightarrow$ outcome	Familiar tool
ALMP	training $\rightarrow$ re-employment $\rightarrow$ earnings	LATE, Heckman selection
Social assistance	reform $\rightarrow$ take-up $\rightarrow$ household income	take-up models
Migration	inflow $\rightarrow$ labour supply $\rightarrow$ native wages	shift-share IV
Returns to schooling	education $\rightarrow$ occupation / skills $\rightarrow$ wages	Mincer decomposition
Wage gap decomp.	group $\rightarrow$ characteristics $\rightarrow$ wage	Oaxaca–Blinder
RCT analysis	treatment arm $\rightarrow$ behaviour $\rightarrow$ outcome	IV, control function

> Existing tools need a functional form (Oaxaca–Blinder) or an instrument for M.
> We ask a smaller question: *is the full-mediation story consistent with the data?*

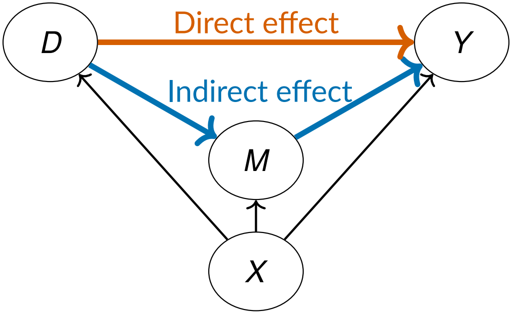
Why mechanisms matter

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

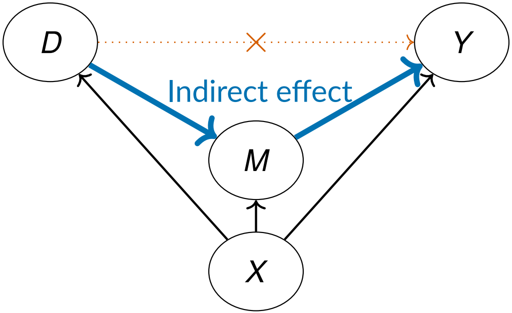
The identification challenge

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

Decomposition of effects



Full mediation condition



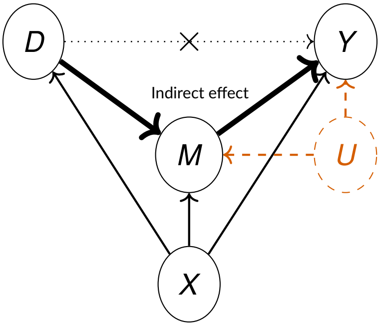
Why testing full mediation is hard

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

Problem: if M is endogenous, conditioning on it opens a collider ($\mathrm{D} \rightarrow \mathrm{M} \leftarrow \mathrm{U} \rightarrow \mathrm{Y}$)

Bias arises *even when D is randomized*

M is a textbook *bad control* ▸ Algebra



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

How it sits next to IV and sensitivity analysis

▸ Map

Contribution

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

> The test needs no instrument for M.

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	$M \perp\!\!\!\perp Y(d,m) \mid D, X$

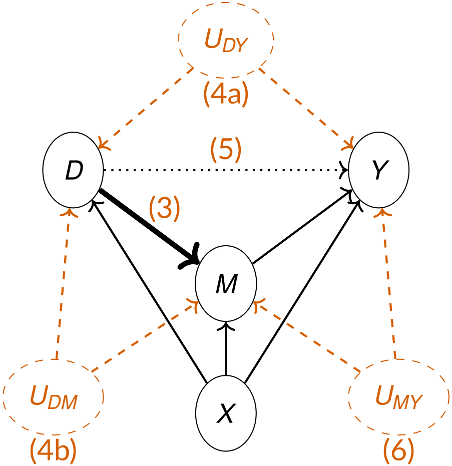
Identifying Assumptions

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)

$$\boxed{Y \perp\!\!\!\perp D \mid M, X}$$

Outcome independent of treatment,
conditional on mediator and covariates

What is the test really doing?

Given the mediator M and covariates X, **does D still help predict Y?**

Reject (TI)

- a direct path $D \to Y$, **or**
- M is endogenous, **or both**

Do not reject (TI)

- consistent with both holding
- not proof of either (power)

> **Job training.** Rejection: knowing re-employment and X, the training arm still predicts long-run earnings.

Theorem 1 (randomly assigned treatment)

> If treatment is randomly assigned and it does move the mediator:
>
> full mediation **and** exogenous mediator $\iff$ $Y \perp\!\!\!\perp D \mid M, X$

In assumption numbers: under 1, 3, 4 we have 5, 6 $\iff$ (TI)

- One test covers **both** conditions at once
- Rejection means at least one of them fails
- Proved over all graphs on (X, D, M, Y) with pairwise confounders

Graphs

Theorem 2 (observational data)

> Without randomisation, if treatment still moves the mediator:
>
> no direct D–Y confounding, full mediation, exogenous mediator
>
> $$\iff \quad Y \perp\!\!\!\perp D \mid M, X$$

In assumption numbers: under 1, 3 we have 4a, 5, 6 $\iff$ (TI)

- Both stay testable without an RCT
- Treatment–mediator confounding (4b) is **not** detected by the test
- Non-rejection alone does not identify the indirect effect

Relation to front-door / back-door

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

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

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

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

Implementation

Estimation

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

- Estimated by **double machine learning**
- Lasso or random forests for the nuisances, so X can be high-dimensional
- Doubly robust, Neyman-orthogonal, 5-fold cross-fitting
- $\sqrt{n}$-consistent and asymptotically normal

Score function ▸ Score · why mean independence ▸ Why

Simulation design

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

$$M = 0.5\,D + X'\beta + \delta\,U_1 + U_2$$

$$Y = M + X'\beta + \gamma D + \delta\,U_1 + U_3$$

	Parameters	Violation tested
Null holds	$\delta = 0,\ \gamma = 0$	none
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**

Size is unaffected by D–M confounding · power is limited at $n = 1{,}000$

Using the test in practice

What you need

- D & binary treatment
- M & the claimed mediator
- X & covariates, can be many

What it will not do

- tell you *which* assumption failed
- detect D–M confounding
- give the size of the direct effect

On sample size. A moderate direct effect ($\gamma = 0.2$) is caught about half the time at $n = 1{,}000$, almost always at $n = 4{,}000$. In small samples a non-rejection is weak evidence.

Empirical Illustrations

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

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

Controls: baseline beliefs, employment, education, demographics · $n = 284$

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

The stronger rejection may also reflect mediator endogeneity

More

Perinatal depression (Baranov et al. 2020, Pakistan)

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, $\ldots \cdot n \sim 600$

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

▸ More

Conclusion

Summary

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

- **Randomized** D**:** full mediation *and* identifiability 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

Is there a specific channel through which the treatment works dominant?
TI can be used to test it.

Thank you for your attention.

`lukaslaffers.github.io`

Double Machine Learning (score)

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

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

Back

The bad-control problem, in algebra

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

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

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 M opens the collider.

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

Back

Angrist–Pischke would say…

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

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

How this fits with other approaches

Approach	Key requirement	What it delivers
IV for M	valid IV $Z \to M$	point ID of indirect effect
Control function	model of M + error distribution	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 the effects
Partial ID (Kwon–Roth)	monotonicity, finite M	sharp bounds and test
Our test (TI)	randomised D + Asm. 1, 3	joint test of both; DML, high-dim X

> **Complementary, not competing.** Run the test first. If you reject, IV or sensitivity analysis tells you how far to trust an indirect-effect estimate.

Back

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. forthcoming in *The Review of Economic Studies*

Back to Literature

Why mean independence?

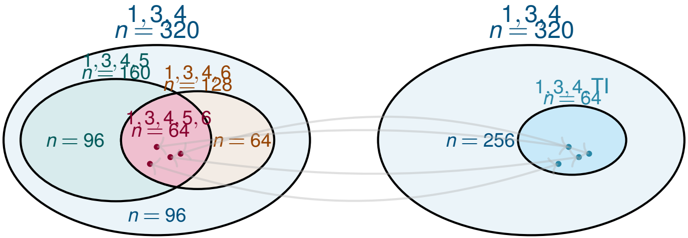
$H_0: \ E[Y \mid M, X, D{=}1] - E[Y \mid M, X, D{=}0] = 0$

Why mean independence instead of full independence?

- **ATE**: mean independence suffices
- **full CI**: needs non-parametric distributional tests
- **mean independence**: allows doubly robust DML

Back to Testing

Theorem 1 (graph classes)

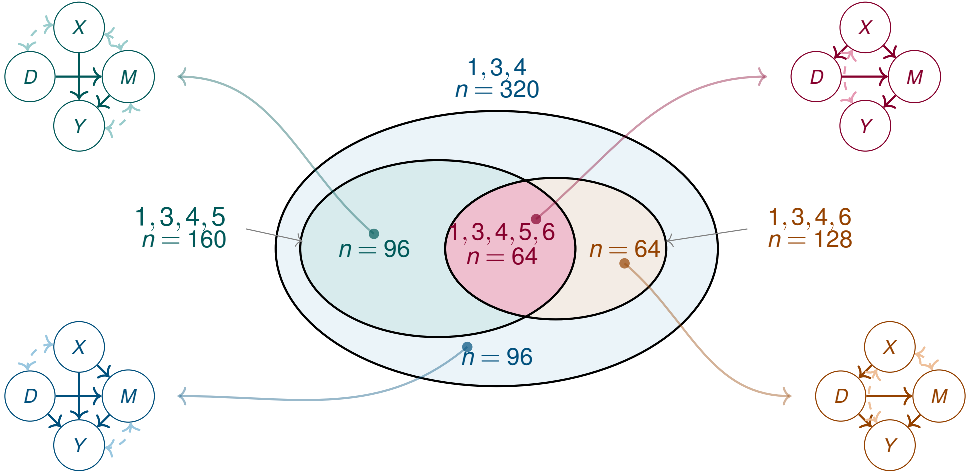


Under Assumptions 1, 3, 4 :

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

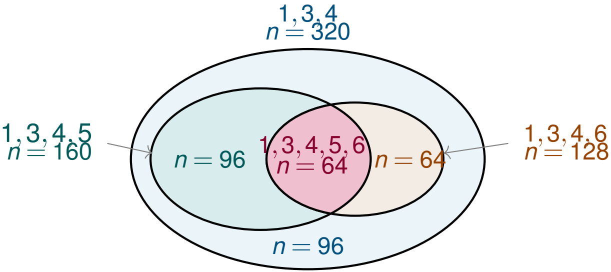
▸ More

Theorem 1 (visual)



Graph search (counting)

- 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

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.

Back

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.

Back

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. forthcoming in *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.