

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

Uni Lausanne 2026

Mediation analysis

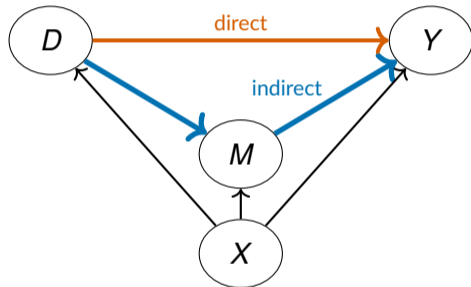
A treatment D affects an outcome Y ...

...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 $D \rightarrow Y$



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

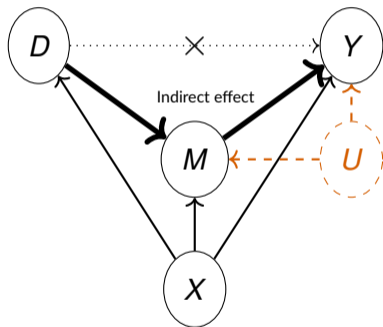
- ▶ M is rarely randomised
- ▶ unobserved M - Y confounders bias estimates
- ▶ hard to test even with randomised D

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



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 and Notation

Observed variables

$D \in \{0, 1\}$ treatment

M mediator

Y outcome

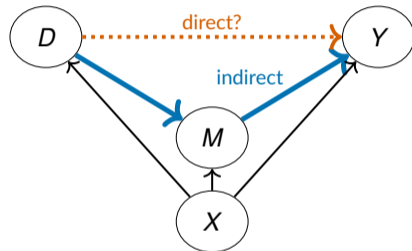
X covariates

Potential outcomes

$M(d)$ mediator under $D=d$

$Y(d, m)$ outcome under $(D, M)=(d, m)$

$Y(m)$ outcome: full mediation



Full mediation: $Y(d, m) = Y(m)$ a.s.
 \Leftrightarrow no direct path $D \rightarrow Y$ bypassing M

Identifying Assumptions

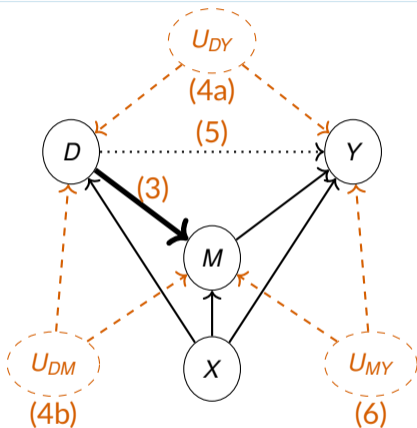
Maintained

- 1 causal ordering + faithfulness (see below)
- 2 common support: $p(D, M | X) \in (0, 1)$
- 3 first stage: $\Pr(M(1) \neq M(0)) > 0$

Under test

- 4 random treatment:
 $D \perp\!\!\!\perp Y(d, m), M(d) | X$
- 5 full mediation: $Y(d, m) = Y(m) | X$ a.s.
- 6 exog. mediator: $M \perp\!\!\!\perp Y(m) | X$

Assumption 1 (causal ordering): allowed directed edges among $\{X, D, M, Y\}$ are $X \rightarrow D, X \rightarrow M, X \rightarrow Y, D \rightarrow M, D \rightarrow Y, M \rightarrow Y$ only. Plus faithfulness.



The key testable implication (TI)

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

Outcome independent of treatment,
conditional on mediator and covariates

Theorem 1 – randomly assigned treatment

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



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

What is the test really doing?

Once we know (M, X) , does treatment assignment D remain predictive of Y ?

Reject TI

D predicts Y given (M, X) . At least one of:

- ▶ direct path $D \rightarrow Y$ exists (A5 fails)
- ▶ M - Y latent confounder (A6 fails)
- ▶ both

Cannot separate the two causes, but the full-mediation story is untenable either way.

Do not reject TI

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

- ▶ does *not* prove either (power)
- ▶ but the data is compatible with the story

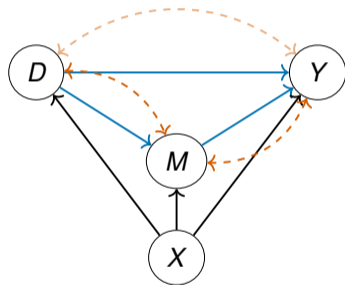
Graph Representation

Graph Representation: Directed + Bidirected Edges

A **directed mixed graph** $\mathcal{G} = (V, E_{\rightarrow}, E_{\leftrightarrow})$:

- ▶ $V = \{X, D, M, Y\}$ observed variables
- ▶ $E_{\rightarrow} \subseteq V \times V$ directed edges, (E_{\rightarrow}, V) acyclic (a DAG)
- ▶ $E_{\leftrightarrow} \subseteq \binom{V}{2}$ bidirected edges, each $V_i \leftrightarrow V_j$ represents a latent common cause

Example – general (no assumptions yet):



Blue = directed edges E_{\rightarrow}

Red dashed = bidirected edges E_{\leftrightarrow} (latent confounders)

Paths and colliders

A **path** in \mathcal{G} is a sequence of adjacent edges.

Node C is a **collider** on path π iff *both* adjacent edges carry an arrowhead at C :

$$\dots \rightarrow C \leftarrow \dots, \quad \dots \rightarrow C \leftrightarrow \dots, \quad \dots \leftrightarrow C \leftrightarrow \dots$$

Otherwise C is a **non-collider**.

Active path given Z

Path π is **active** given $Z \subseteq V$ iff:

1. every *non-collider* on π is $\notin Z$
2. every *collider* on π is an ancestor of a node in Z

$A \perp_{\mathcal{G}} B \mid Z$: no active path from A to B given Z

Global Markov: $A \perp_{\mathcal{G}} B \mid Z \Rightarrow A \perp\!\!\!\perp B \mid Z$
Faithfulness (Asm 1): converse holds

Key example – why TI can fail

Suppose $D \rightarrow Y$ is absent but $M \leftrightarrow Y$ is present.
Is there an active path from D to Y given $Z = \{M, X\}$?



collider on π

Path π : $D \rightarrow M \leftrightarrow Y$

M is a collider: $\rightarrow M \leftrightarrow$

$M \in Z = \{M, X\} \Rightarrow$ collider condition \checkmark

Path is **active** given $\{M, X\}$!

Conditioning on M opens the collider path.
 $D \not\perp_{\mathcal{G}} Y \mid M, X$ even without $D \rightarrow Y$.
 \Rightarrow TI fails if $M \leftrightarrow Y$ is present.

Enumerating All Graphs over $V = \{X, D, M, Y\}$

Directed part: DAGs on 4 labeled nodes

Number of DAGs on n labeled nodes (OEIS A003024):

n	1	2	3	4
#DAGs	1	3	25	543

Bidirected part: latent confounders

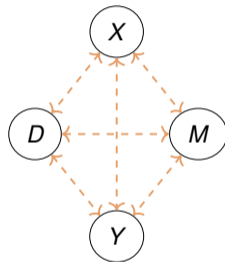
Possible pairs: $\binom{4}{2} = 6$

Each pair: present or absent

Subsets: $2^6 = 64$

$543 \times 64 = \mathbf{34,752}$
distinct graphs over $\{X, D, M, Y\}$

The 6 possible bidirected edges:



Each independently present/absent.

Assumption 1: Causal Ordering and Faithfulness

Exclusion restrictions (causal ordering): Assumption 1 fixes a partial order $X \prec D, M, Y$ and $D \prec M, Y$ with $M \prec Y$, forbidding reverse-direction edges. Allowed directed edges:

$$\{X \rightarrow D, X \rightarrow M, X \rightarrow Y, D \rightarrow M, D \rightarrow Y, M \rightarrow Y\}$$

Any subset of these 6 is acyclic. All 6 bidirected edges (latent confounders) remain possible.

Faithfulness (A1): conditional independence in distribution \Leftrightarrow d-separation in graph.

Component	Count
Directed (any subset of 6)	$2^6 = 64$
Bidirected (any subset of 6)	$2^6 = 64$
Satisfying A1	4,096 = 64×64

Assumptions 3 and 4: Filtering to 320 Graphs

Assumption 3: $D \rightarrow M$ exists

Requires $D \rightarrow M \in E_{\rightarrow}$. Simply halves the count.

Satisfying Asm 1 4,096

+ Asm 3 2,048

Assumption 4: Treatment exogeneity

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

- ▶ 4a: $D \perp\!\!\!\perp Y(d, m) \mid X$ – no D - Y confounding given X (eliminates $D \leftrightarrow Y$)
- ▶ 4b: $D \perp\!\!\!\perp M(d) \mid X$ – no D - M confounding given X (eliminates $D \leftrightarrow M$)

In graph terms: d -separation of D from $\{M, Y\}$ given X in the interventional graph (remove outgoing D edges).

Verified computationally by checking d -separation in each of the 2,048 graphs:

+ Asm 4 320

1,728 graphs eliminated by Asm 4. **320 survive:** our working universe for Theorem 1.

From Potential Outcomes to m-Separation: SWIGs

PO \leftrightarrow graph dictionary:

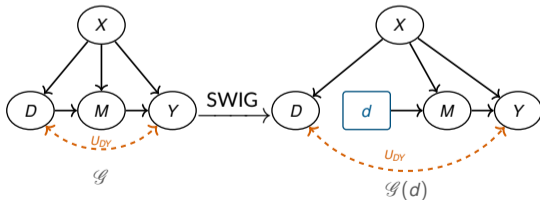
A4: $D \perp\!\!\!\perp \{Y(d, m), M(d)\} \mid X \iff D \perp\!\!\!\perp_{\mathcal{G}(d)} \{Y, M\} \mid X$

A5: no direct effect $\iff D \rightarrow Y \notin E_{\rightarrow}$

A6: $M \perp\!\!\!\perp Y(m) \mid X \iff M \perp\!\!\!\perp_{\mathcal{G}(m)} Y \mid X$

TI: $Y \perp\!\!\!\perp D \mid M, X \iff Y \perp\!\!\!\perp_{\mathcal{G}} D \mid M, X$

SWIG (Richardson & Robins 2013)



Theorem 1 – Proof Sketch

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



(\implies) Assumptions imply TI

Asm 5 removes $D \rightarrow Y$; Asm 6 removes active $M \rightarrow Y$ paths.

After Asm 4, every path $D \rightsquigarrow Y$ goes through M .

Conditioning on M blocks $D \rightarrow M \rightarrow Y$ (non-collider) and Asm 4 kills all collider bypasses.

\implies no active path $D \rightsquigarrow Y$ given $\{M, X\}$.

(\impliedby) TI implies Assumptions

By faithfulness: $\text{TI} \implies D \perp_{\mathcal{G}} Y \mid M, X$.

If Asm 5 failed: $D \rightarrow Y$ active given $\{M, X\}$ – contradiction.

If Asm 6 failed: path $D \rightarrow M \leftrightarrow Y$ active (collider $M \in Z$) – contradiction.

\implies Both Asm 5 and 6 must hold.

Assumptions 5 & 6 vs. Testable Implication

Among the **320 graphs** satisfying $\{1,3,4\}$:

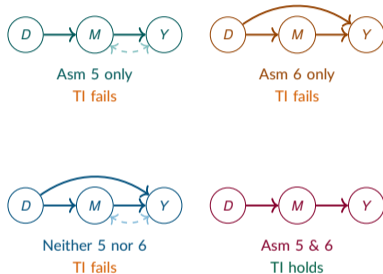
Condition	#Graphs
Asm 5 alone ($D \rightarrow Y \notin E_{\rightarrow}$)	160
Asm 6 alone (no active $M \rightarrow Y$ path given X)	128
Asm 5 and Asm 6	64
TI: $Y \perp_{\mathcal{G}} D \mid M, X$ (d-separation check)	64

Verification by enumeration:

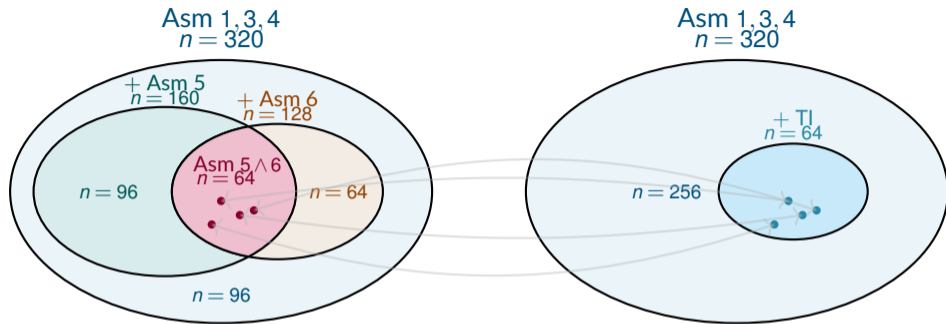
	TI false	TI true
$A5 \wedge A6$ false	256	0
$A5 \wedge A6$ true	0	64

Both off-diagonals empty \Rightarrow biconditional holds over all 320 graphs.

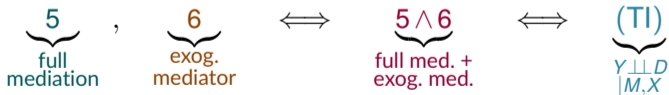
Example graphs from each region:



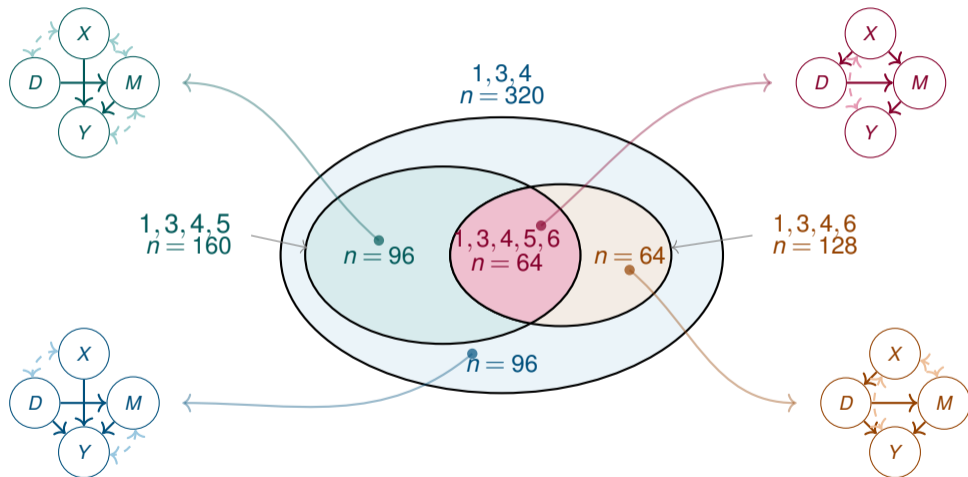
(X omitted for clarity; TI status via d-separation)



Under Assumptions 1, 3, 4 :



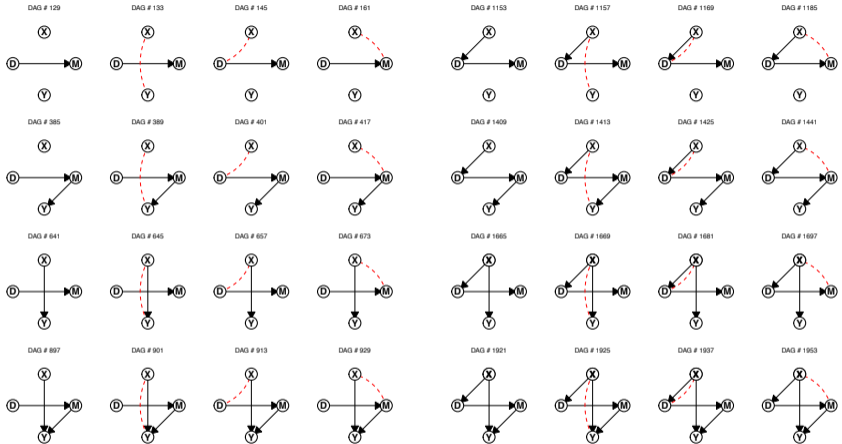
Theorem 1 – All DAG classes (visual)



Graph search – counting details

- ▶ $543 \times 64 = 34,752$ causal graphs over $\{X, D, M, Y\}$ with pairwise latent confounders
- ▶ 4,096 satisfying Assumption 1 (causal ordering + faithfulness)
- ▶ 2,048 additionally satisfying Assumption 3 ($D \rightarrow M$ present)
- ▶ 320 additionally satisfying Assumption 4 (treatment exogeneity)
- ▶ 64 additionally satisfying Assumptions 5 and 6 = exactly those satisfying TI

Each implication ($A \perp_{\mathcal{G}} B \mid Z?$) checked by the d-separation algorithm. The computation confirms the theorem for all graphs in the class – a complete case analysis over a finite, exhaustively enumerated model space.



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



Theorem 2 – Observational Data

Theorem 2. Under Assumptions 1, 3 only (no treatment randomization assumed):

$$\underbrace{4a}_{\text{no } D\text{-}Y \text{ confounding}}, \quad \underbrace{5}_{\text{full mediation}}, \quad \underbrace{6}_{\text{exog. mediator}} \iff \underbrace{(TI)}_{Y \perp\!\!\!\perp D \mid M, X}$$

- ▶ Full mediation (Asm 5) and mediator exogeneity (Asm 6) remain jointly testable in observational settings
- ▶ **Treatment-mediator confounding** (Asm 4b fails) is *not* detected by TI – it enters symmetrically on both sides
- ▶ Rejection \Rightarrow at least one of $\{4a, 5, 6\}$ fails

Theorems 3 & 4 – Front-door / Back-door

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

Theorem 3. (TI) \Rightarrow (BD = FD) **Theorem 4.** (BD = FD) + **Separability** \Rightarrow (TI)

- ▶ BD=FD is *weaker* than TI: implied by TI, but converse needs extra structure
- ▶ Separability ($\Pr(Y=y \mid D, M, X) = \alpha(y, D, X) + \beta(y, M, X)$) rules out D - M interactions – strong assumption
- ▶ BD=FD: conservative diagnostic (rejection \Rightarrow TI fails; non-rejection $\not\Rightarrow$ TI)
- ▶ Pearl's front-door criterion *assumes* full mediation; our test *checks* it

Statistical Implementation

Double Machine Learning

$$H_0: E[Y | M, X, D=1] - E[Y | M, X, D=0] = 0$$

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

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

$$\mu_d = E[Y | M, X, D=d], \quad \rho = \Pr(D=1 | M, X)$$

- ▶ **Neyman orthogonality:** $E[\partial_\eta \tilde{\psi}] = 0$ – score insensitive to first-order nuisance error
- ▶ **5-fold cross-fitting:** avoids overfitting bias, maintains efficiency
- ▶ **Lasso / ML nuisances:** \sqrt{n} -consistent, asymptotically normal test statistic
- ▶ **Mean independence H_0 :** weaker than full $Y \perp\!\!\!\perp D | M, X$; enables DML with high-dim X

Why mean independence?

$$H_0 : E[Y | M, X, D=1] - E[Y | M, X, D=0] = 0$$

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

Simulations

Simulation design

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

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

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

	Parameters	Violation tested
Null holds	$\delta = 0, \gamma = 0$	—
Null fails	$\delta \neq 0$	$D\text{-}M\text{-}Y$ confounding
Null fails	$\gamma \neq 0$	direct effect $D \rightarrow Y$

$p = 200$ covariates · $n \in \{1000, 4000\}$ · 1000 replications · lasso, 5-fold cross-fitting

Theorem 1: $\lambda = 0$ · Theorem 2: $\lambda = 0.25$ (introduces $D\text{-}M$ confounding)

Results – rejection rates at the 5% level

	Scenario	$n = 1,000$	$n = 4,000$
<i>Main design, $\lambda = 0$</i>			
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
<i>With D-M confounding, $\lambda = 0.25$</i>			
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 other men are about women working outside the home
- ▶ Experiment: receive information on peers' beliefs, or not
- ▶ Offered job-search service sign-up or gift card
- ▶ Treated men's wives more likely to apply for job interviews within 6 months
- ▶ Is short-run sign-up for job-search service the main driver of long-run outcomes?
- ▶ $n = 284$

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	<i>p</i> -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 long-run effect on financial empowerment
- ▶ Grandmother presence (family support) and relationship with husband as mediators
- ▶ $n \approx 600$

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	<i>p</i> -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*

Conclusion

Summary

$Y \perp\!\!\!\perp D \mid M, X$ jointly tests **full mediation (Asm 5)** and **mediator exogeneity (Asm 6)**

- ▶ **Graph-theoretic proof:** equivalence over all 320 graphs satisfying $\{1, 3, 4\}$
- ▶ **Randomized D (Thm 1):** $\text{Asm 5} \wedge \text{Asm 6} \iff \text{TI}$
- ▶ **Observational D (Thm 2):** $\text{Asm 4a} \wedge 5 \wedge 6 \iff \text{TI}$; D - M confounding not detected
- ▶ **Front-door (Thm 3&4):** $\text{BD}=\text{FD} \Leftarrow \text{TI}$; $\text{BD}=\text{FD} \Rightarrow \text{TI}$ only under separability
- ▶ **DML:** \sqrt{n} -consistent, doubly robust, Neyman-orthogonal, high-dim X
- ▶ Both applications **reject** – mediator endogeneity likely present

Is there a specific channel through which the treatment works dominant?

TI can be used to test it.

Comparison to Kwon and Roth (2026)

Kwon & Roth (2026)

- ▶ Sharp null of full mediation
- ▶ Partial identification via linear program
- ▶ Sharp bounds on degree of violation
- ▶ Finite-support mediator required
- ▶ Monotonicity assumptions
- ▶ Robust to many covariates X

Our test (TI)

- ▶ Joint test: full mediation *and* mediator exogeneity
- ▶ Continuous or discrete M
- ▶ No monotonicity
- ▶ DML – high-dimensional X
- ▶ Sharper in applications (picks up mediator endogeneity)

Key difference. KR tests full mediation alone (treats mediator as exogenous). We test *both* jointly – rejection may reflect direct effects *or* mediator endogeneity. The extra power in our applications comes precisely from detecting the latter.