

# 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

EuroCIM Oxford 2026

A joint test of

**full mediation**

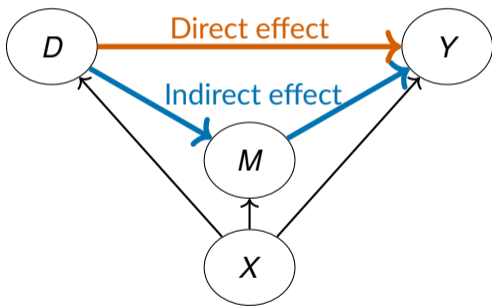
and

**identifiability of causal mechanisms**

in mediation analysis

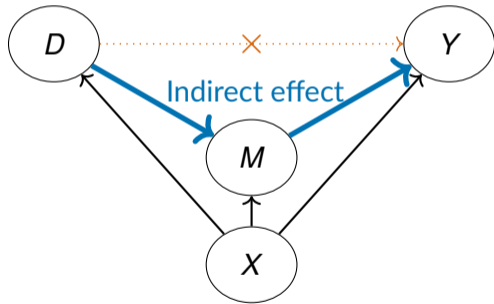
## Decomposition of effects

---



## Full mediation condition

---

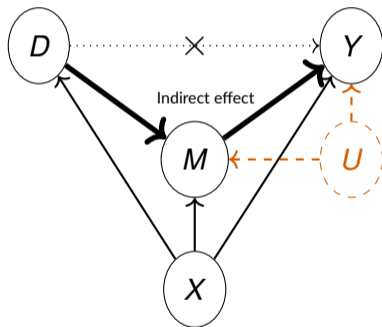


## Why testing full mediation is hard

**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



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

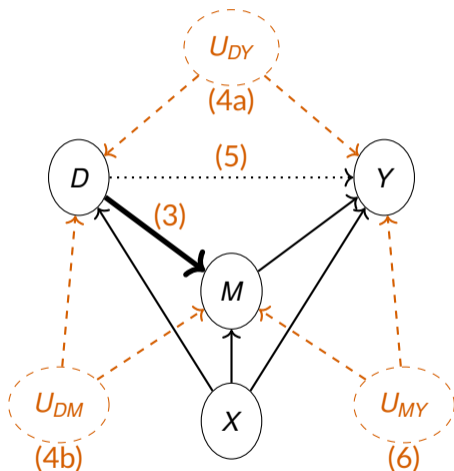
# Identifying Assumptions

## Assumption 1

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

## Assumption 3

- ▶  $D \rightarrow M$  exists



The key **testable implication (TI)**

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

Outcome independent of treatment,  
conditional on mediator and covariates

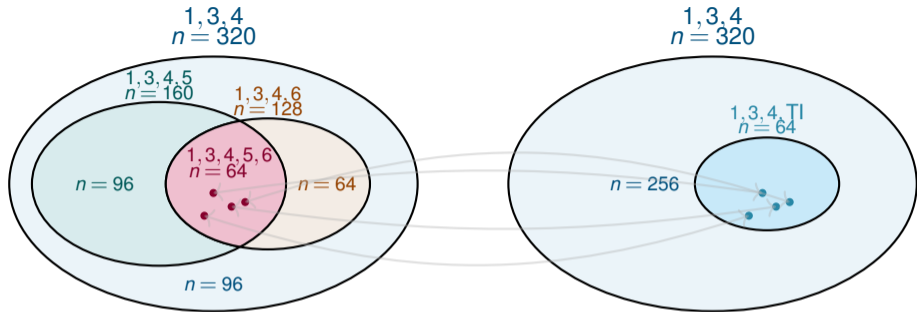
## Theorem 1 – randomly assigned treatment

---

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



Under Assumptions 1, 3, 4 :



► More

## Theorem 2 – observational data

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



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

## Relation to front-door / back-door

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

**Theorem 3.** (TI)  $\Rightarrow$  (BD = FD)

**Theorem 4.** (BD = 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

---

# Implementation

---

# Double Machine Learning

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

▶ More

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

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

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

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

---

# Simulation

---

## Simulation design

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

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

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

	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, <math>\lambda = 0</math></i>			
Null	$\delta = 0, \gamma = 0$	0.055	0.040
M-Y conf.	$\delta = 0.25, \gamma = 0$	0.969	<b>1.000</b>
Direct effect	$\delta = 0, \gamma = 0.2$	0.532	<b>1.000</b>
<i>With D-M confounding, <math>\lambda = 0.25</math></i>			
Null	$\delta = 0, \gamma = 0$	0.030	0.037
M-Y conf.	$\delta = 1, \gamma = 0$	<b>1.000</b>	<b>1.000</b>
Direct effect	$\delta = 0, \gamma = 0.2$	0.482	<b>0.993</b>

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

---

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

Controls: baseline beliefs, employment, education, demographics

▶ More

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

- 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,...

▶ More

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** 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. 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.

## 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?

---

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

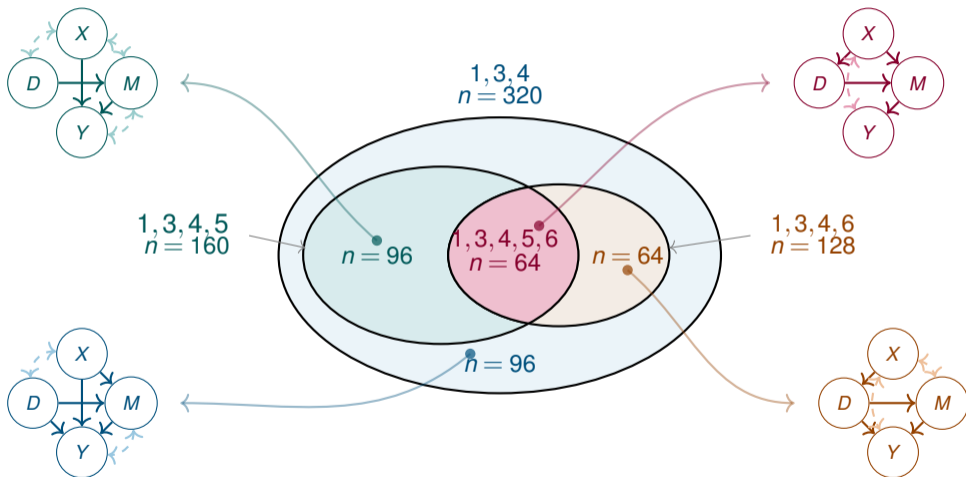
Why mean independence instead of full independence?

It is practical:

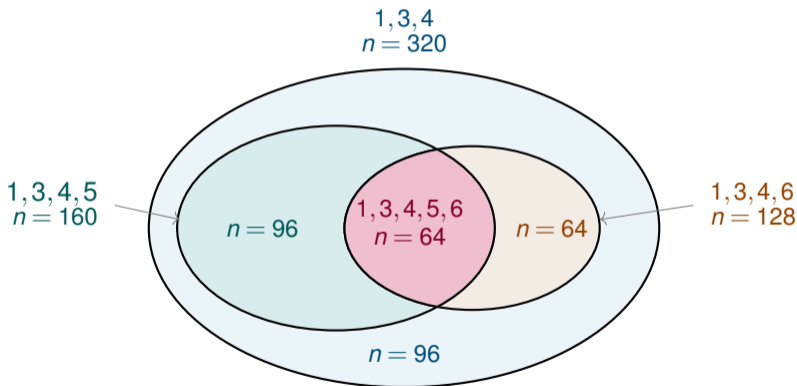
- ▶ **ATE** – for average TE, it suffice to use mean independence
- ▶ **full CI** – requires non-parametric distributional tests - more demanding
- ▶ **mean independence** – allows doubly robust DML framework

▶ [Back to Testing](#)

# Theorem 1 – visual



- ▶ 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



- ▶ 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.

- ▶ RCT of CBT to reduce depression for pregnant women in rural Pakistan
- ▶ studied effect on financial empowerment
- ▶ 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.