Appendix

A Proofs for Section 3

Proof of Proposition 1. We establish the two claims in turn. Throughout, $V \not\perp\!\!\!\perp W \mid U$ is interpreted in the sense of the Convention introduced in Section 2.

Part (i): Conditional Exogeneity, $Z \perp\!\!\!\perp Y(t,m) \mid (T,X)$. Fix $(t,m) \in \mathcal{T} \times \mathcal{M}$ and $z \in \mathcal{Z}$. By marginalization of the joint conditions, IV exogeneity (7) gives $Z \perp\!\!\!\perp (Y(t,m), T(z)) \mid X$, and MCA (14) gives $Y(t,m) \perp\!\!\!\perp T(z) \mid X$. Combining the two factorizations,

$$f(Z,\, Y(t,m),\, T(z) \mid X) \;=\; f(Z \mid X) \cdot f(Y(t,m) \mid X) \cdot f(T(z) \mid X),$$

so Z, $Y(t,m)$, and $T(z)$ are mutually independent given X; in particular, $Y(t,m) \perp\!\!\!\perp (Z, T(z)) \mid X$.

By consistency, $T = T(Z)$, so the conditioning event $\{Z = z,\, T = t'\}$ equals $\{Z = z,\, T(z) = t'\}$. Combined with the displayed mutual independence,

$$f\big(Y(t,m) \,\big|\, Z = z,\, T = t',\, X\big) \;=\; f\big(Y(t,m) \,\big|\, Z = z,\, T(z) = t',\, X\big) \;=\; f\big(Y(t,m) \,\big|\, X\big)$$

for every (z,t') in the support of $(Z,T) \mid X$. The right side does not depend on z, establishing $Z \perp\!\!\!\perp Y(t,m) \mid (T,X)$. The argument in fact establishes the stronger latent-variable exogeneity $Y(t,m) \perp\!\!\!\perp (Z,T) \mid X$; the form stated in the proposition suffices for the moment-based identification arguments of Section 4.

Part (ii): Conditional Relevance, $Z \not\perp\!\!\!\perp M \mid (T,X)$. We invoke the substantive structural condition—maintained throughout the paper—that treatment and mediator share common unobserved determinants, formally

$$\exists\, t \in \mathcal{T}: \quad M(t) \;\not\perp\!\!\!\perp\; \{T(z)\}_{z \in \mathcal{Z}} \,\big|\, X, \tag{95}$$

which in the linear model of Section 4 is equivalent to $\operatorname{cov}(\varepsilon_T, \varepsilon_M \mid X) \neq 0$. Condition (95) is what distinguishes the mediation problem from a setting with two independent IV equations; it is the source of the collider channel illustrated by the dashed link $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_M$ in Table 2.

By consistency $M = M(T)$ and the potential-outcome representation $T = T(Z)$, the conditioning event $\{Z = z, T = t\}$ equals $\{Z = z, T(z) = t\}$, hence

$$\mathcal{L}\big(M \,\big|\, Z = z,\, T = t,\, X\big) \;=\; \mathcal{L}\big(M(t) \,\big|\, Z = z,\, T(z) = t,\, X\big).$$

By marginalization of IV exogeneity (7), $Z \perp\!\!\!\perp (M(t), T(z)) \mid X$, and the graphoid weak-union axiom yields $M(t) \perp\!\!\!\perp Z \mid (T(z), X)$. Therefore

$$\mathcal{L}\big(M \,\big|\, Z = z,\, T = t,\, X\big) \;=\; \mathcal{L}\big(M(t) \,\big|\, T(z) = t,\, X\big). \tag{96}$$

The right-hand side of (96) is a conditional law of $M(t)$ given the latent treatment-response event $\{T(z) = t\}$. Under (95), $M(t)$ depends non-trivially on the latent schedule $\{T(z')\}_{z' \in \mathcal{Z}}$, so the events $\{T(z) = t\}$ and $\{T(z') = t\}$—which select different subpopulations of the latent treatment errors as z varies—generically induce different conditional distributions for $M(t)$. Formally, there exist $z \neq z'$ in

$\mathcal{Z}$ and a positive-measure subset of X-values on which

$$\mathcal{L}\big(M(t)\,\big|\,T(z)=t,\,X\big)\;\neq\;\mathcal{L}\big(M(t)\,\big|\,T(z')=t,\,X\big).$$

Substituting into (96) delivers $\mathcal{L}(M \mid Z=z, T=t, X) \neq \mathcal{L}(M \mid Z=z', T=t, X)$, which by the Convention of Section 2 is $Z \not\perp\!\!\!\perp M \mid (T, X)$. □

Remark. The two parts have distinct roles. Part (i) uses MCA (14) to close the latent treatment–outcome confounding channel and deliver conditional exogeneity. Part (ii) uses only the IV conditions together with the common-drivers condition (95); MCA is *not* invoked for conditional relevance, which arises from the collider mechanism on the path $Z \to T \leftarrow \mathbf{V}_{TM} \to M$. The explicit algebraic verification under linearity appears in equation (20).

Proof of Proposition 2. Conditional on X and for any $(t, m, z) \in \mathcal{T} \times \mathcal{M} \times \mathcal{Z}$, the structural equations (9) and (11) give

$$Y(t,m)\;=\;\big(\alpha+\tau\,t+\theta\,m+X\beta\big)+\varepsilon_Y,\qquad T(z)\;=\;\big(\pi_0+\pi_1\,z+X\pi_2\big)+\varepsilon_T,$$

so $Y(t,m) \mid X$ is an affine bijection of ε_Y (with the affine shift parametrized by (t, m, X)), and $T(z) \mid X$ is an affine bijection of ε_T. The same bijections extend to the entire collections: $\{Y(t,m)\}_{(t,m)} \mid X$ is a deterministic function of ε_Y, and $\{T(z)\}_z \mid X$ is a deterministic function of ε_T.

($\Leftarrow$) If $\varepsilon_Y \perp\!\!\!\perp \varepsilon_T \mid X$, then any deterministic functions of these errors given X are independent given X, giving $\{Y(t,m)\}_{(t,m)} \perp\!\!\!\perp \{T(z)\}_z \mid X$.

($\Rightarrow$) If $\{Y(t,m)\}_{(t,m)} \perp\!\!\!\perp \{T(z)\}_z \mid X$, then for any fixed (t, m, z), $\varepsilon_Y = Y(t,m) - \alpha - \tau t - \theta m - X\beta$ is a measurable function of $Y(t,m)$ given X, and $\varepsilon_T = T(z) - \pi_0 - \pi_1 z - X\pi_2$ is a measurable function of $T(z)$ given X. Hence $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y \mid X$. □

Remark. This equivalence justifies the interchangeable use of MCA (14) (potential-outcome form) and $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y \mid X$ (structural-error form) throughout the paper.

B Unconditional Independence Restrictions

This appendix classifies the three pairwise unconditional independence restrictions on the structural errors $(\varepsilon_T, \varepsilon_M, \varepsilon_Y)$ and establishes that mediated confounding is the unique marginal restriction that identifies the mediation parameters while preserving the endogeneity structure of the IV model. Appendix C extends the analysis to the three conditional restrictions.

The structural mediation model of Table 1 generates counterfactual variables $T(z) = f_T(z, \varepsilon_T)$, $M(t) = f_M(t, \varepsilon_M)$, and $Y(t, m) = f_Y(t, m, \varepsilon_Y)$, where each counterfactual at fixed arguments depends on a single error term. Joint instrument exogeneity, $Z \perp\!\!\!\perp (\varepsilon_T, \varepsilon_M, \varepsilon_Y)$, combined with arbitrary dependence among the errors, yields the following independence structure:

(i) *Instrument exogeneity*: $Z \perp\!\!\!\perp T(z)$, $Z \perp\!\!\!\perp M(t)$, and $Z \perp\!\!\!\perp Y(t, m)$ for all (t, m, z).

(ii) *Reduced-form exogeneity*: $Z \perp\!\!\!\perp M(z)$ and $Z \perp\!\!\!\perp Y(t)$, where $M(z) = f_M(T(z), \varepsilon_M)$ and $Y(t) = f_Y(t, M(t), \varepsilon_Y)$.

(iii) *Treatment endogeneity*: $T \not\perp\!\!\!\perp M(t)$ and $T \not\perp\!\!\!\perp Y(t)$—both M and Y are endogenous with respect to T.

(iv) *Joint endogeneity of* (T, M): $Y(t, m) \not\perp\!\!\!\perp (T, M)$—the outcome at fixed treatment and mediator depends on both endogenous variables.

(v) *No unconditional instrument for* $M \to Y$: $Y(m) \not\perp\!\!\!\perp Z$—the instrument is not exogenous for Y when only the mediator argument is fixed, because $Y(m) = f_Y(T, m, \varepsilon_Y)$ retains dependence on T and hence on Z through $T(z)$.

Properties (i)–(ii) ensure that Z identifies the total effect of T on Y and the effect of T on M. Properties (iii)–(v) establish that, without further restrictions, the instrument cannot disentangle the direct effect from the mediator effect (Section 2).

B.1 Uniqueness Among Marginal Restrictions

The following proposition establishes that, among the three marginal (unconditional) pairwise independence restrictions, MCA is the unique one that generates a valid conditional instrument for the mediator while preserving the endogeneity properties of the IV mediation model.

Proposition 7 (Uniqueness Among Marginal Restrictions)**.** *Consider the structural mediation model of Table 1 under the mediation IV conditions (6)–(8). A pairwise independence restriction on* $(\varepsilon_T, \varepsilon_M, \varepsilon_Y)$ enables mediation identification *if it simultaneously delivers:*

(a) ***Conditional exogeneity****:* $Z \perp\!\!\!\perp Y(t, m) \mid (T, X)$ *for all* (t, m)*;*

(b) ***Conditional relevance****:* $Z \not\perp\!\!\!\perp M \mid (T, X)$*;*

(c) ***Treatment–mediator endogeneity****:* $T \not\perp\!\!\!\perp M(t)$*;*

(d) ***Mediator–outcome endogeneity****:* $M \not\perp\!\!\!\perp Y(t, m) \mid T$.

Among the three pairwise independence restrictions—(I) $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y$*, (II)* $\varepsilon_T \perp\!\!\!\perp \varepsilon_M$*, and (III)* $\varepsilon_M \perp\!\!\!\perp \varepsilon_Y$*—only (I) enables mediation identification.*

Proof. All independence statements condition on X, which is suppressed for brevity.

Case (I): $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y$. Since $T(z) = f_T(z, \varepsilon_T)$ and $Y(t, m) = f_Y(t, m, \varepsilon_Y)$, this implies $Y(t, m) \perp\!\!\!\perp T(z)$ for all (t, m, z), which is MCA (14). Proposition 1 then delivers conditional exogeneity (a), and the collider mechanism analyzed in Section 3 delivers conditional relevance (b), provided the treatment and mediator share common unobserved determinants. Restriction (I) places no constraint on $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_M$ or $\varepsilon_M \not\perp\!\!\!\perp \varepsilon_Y$, so (c) and (d) are preserved. All four conditions are satisfied.

Case (II): $\varepsilon_T \perp\!\!\!\perp \varepsilon_M$. Since $T(z) = f_T(z, \varepsilon_T)$ depends only on ε_T and $M(t) = f_M(t, \varepsilon_M)$ depends only on ε_M, this implies $T(z) \perp\!\!\!\perp M(t)$ for all (z, t). Marginalizing over Z: $T \perp\!\!\!\perp M(t)$, violating condition (c). Moreover, the collider mechanism that generates conditional relevance requires $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_M$: conditioning on $T = f_T(Z, \varepsilon_T)$ induces $Z \not\perp\!\!\!\perp \varepsilon_T \mid T$, but when $\varepsilon_T \perp\!\!\!\perp \varepsilon_M$, this dependence does not propagate to ε_M, so $Z \perp\!\!\!\perp M \mid T$. Condition (b) is also violated.

Case (III): $\varepsilon_M \perp\!\!\!\perp \varepsilon_Y$. Since $M(t) = f_M(t, \varepsilon_M)$ and $Y(t, m) = f_Y(t, m, \varepsilon_Y)$, this implies $M(t) \perp\!\!\!\perp Y(t, m)$ for all (t, m), so $M \perp\!\!\!\perp Y(t, m) \mid T$: the mediator is exogenous conditional on treatment. This violates condition (d). When M is exogenous with respect to Y given T, the causal effect of M on Y is identified by regressing Y on (T, M) without instrumenting—the IV mediation problem does not arise.

Since (II) violates (b) and (c) and (III) violates (d), only (I) enables mediation identification. □

C Conditional Independence Restrictions and the Control Function

This appendix investigates the three pairwise conditional independence restrictions on the structural errors $(\varepsilon_T, \varepsilon_M, \varepsilon_Y)$:

$$\text{(IV)}\ \ \varepsilon_T \perp\!\!\!\perp \varepsilon_Y \mid \varepsilon_M, \qquad \text{(V)}\ \ \varepsilon_T \perp\!\!\!\perp \varepsilon_M \mid \varepsilon_Y, \qquad \text{(VI)}\ \ \varepsilon_M \perp\!\!\!\perp \varepsilon_Y \mid \varepsilon_T. \tag{97}$$

We establish two results. First, all three restrictions preserve the endogeneity conditions required for a nontrivial mediation problem (Section C.2). Second, the identification potential of these restrictions is naturally assessed through the control function (CF) approach: restrictions (IV) and (V) fail because the implied CF destroys the variation needed for identification (Section C.3), while restriction (VI) succeeds (Section C.4). Section C.5 explains why the paper nonetheless adopts MCA over restriction (VI).

Within this class, exactly two restrictions satisfy the requirements for mediation identification while preserving the endogenous mediation structure. The first is mediated confounding, $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y$, which identifies the mediator effect through the conditional IV argument developed in Section 3.3. The second is $\varepsilon_M \perp\!\!\!\perp \varepsilon_Y \mid \varepsilon_T$, which identifies the mediator effect through the control-function approach developed below. No other restriction in this class delivers identification while preserving the model's endogeneity structure.

C.1 The Control Function Approach

Consider the triangular system $T = f_T(Z, \varepsilon_T)$, $Y = f_Y(T, \varepsilon_Y)$ with $Z \perp\!\!\!\perp (\varepsilon_T, \varepsilon_Y)$ and $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_Y$. Treatment is endogenous because T depends on ε_T, which covaries with ε_Y. The control function approach restores exogeneity by exploiting the fact that ε_T is a sufficient summary of the confounding between T and Y: the triangular structure and the exclusion of Z from the outcome equation imply

$$T \perp\!\!\!\perp \varepsilon_Y \,\big|\, \varepsilon_T. \tag{98}$$

Since $Y(t) = f_Y(t, \varepsilon_Y)$ depends on ε_Y alone, condition (98) implies $T \perp\!\!\!\perp Y(t) \mid \varepsilon_T$ for all t: conditional on ε_T, treatment is as good as randomly assigned with respect to the potential outcome. This is a *matching condition*—ε_T plays the role of a selection variable that, once conditioned on, restores exogeneity and identifies the causal effect of T on Y.

Two assumptions are required to operationalize (98). First, *invertibility*: the structural function $f_T(z, \cdot)$ must be strictly monotone in ε_T for each z, so that $\varepsilon_T = f_T^{-1}(T, Z)$ is uniquely recoverable from observed data. The control variable $U_T = F_{T|Z}(T, Z)$ is then a monotone transformation of ε_T, uniformly distributed on $[0, 1]$. This requires that the treatment be continuously distributed conditional on Z. Second, *conditional mean independence*: the full conditional independence (98) implies $\mathrm{E}[\varepsilon_Y \mid T, \varepsilon_T] = \mathrm{E}[\varepsilon_Y \mid \varepsilon_T] \equiv h(\varepsilon_T)$, so that T contributes no additional information about ε_Y beyond what ε_T already provides. In the linear model with $T = \pi_1 Z + \varepsilon_T$ and $Y = g(T) + \varepsilon_Y$, this yields the augmented regression $\mathrm{E}[Y \mid T, \varepsilon_T] = g(T) + h(\varepsilon_T)$; the effect of T is identified from the residual variation in T after partialling out ε_T, which is driven by the exogenous instrument.

In the mediation model (9)–(12), there are three error terms and the goal is to identify the separate effects (τ, θ) rather than a single total effect. The question becomes: *which error term should serve as the control function?* Each conditional restriction in (97) corresponds to conditioning on a different error

term—ε_M for (IV), ε_Y for (V), and ε_T for (VI). As we show below, only conditioning on ε_T succeeds.

C.2 Preservation of Endogeneity

All three conditional restrictions preserve the endogeneity conditions required for a nontrivial mediation problem:

$$T \not\perp\!\!\!\perp M(t), \qquad T \not\perp\!\!\!\perp Y(t), \qquad M \not\perp\!\!\!\perp Y(t,m) \mid T. \tag{99}$$

Conditional independence does not imply marginal independence: $\varepsilon_j \perp\!\!\!\perp \varepsilon_k \mid \varepsilon_\ell$ places no constraint on the marginal relationship $\varepsilon_j \not\perp\!\!\!\perp \varepsilon_k$. Each of (IV)–(VI) is therefore compatible with all three marginal dependences $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_M$, $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_Y$, and $\varepsilon_M \not\perp\!\!\!\perp \varepsilon_Y$ holding simultaneously. Conditions (99) are driven by these marginal dependences: $T \not\perp\!\!\!\perp M(t)$ requires $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_M$; $T \not\perp\!\!\!\perp Y(t)$ requires $\varepsilon_T \not\perp\!\!\!\perp (\varepsilon_M, \varepsilon_Y)$; and $M \not\perp\!\!\!\perp Y(t,m) \mid T$ requires $\varepsilon_M \not\perp\!\!\!\perp \varepsilon_Y$.

In particular, since $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_Y$ is compatible with each conditional restriction, none implies the unconditional independence $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y$ that characterizes MCA. The conditional restrictions therefore do not generate the IV exogeneity condition of Proposition 1, nor do they inadvertently solve the endogeneity problem. The model under any of (IV)–(VI) exhibits the same endogeneity as the unrestricted model; only the conditional dependence structure changes, which governs the feasibility of the control function approach.

C.3 Restrictions (IV) and (V): Failure of the Control Function

Restriction (IV): $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y \mid \varepsilon_M$. Under this restriction, the confounding between ε_T and ε_Y is absorbed by conditioning on ε_M. If the mediator equation $f_M(t,\cdot)$ is strictly monotone in ε_M for each t and M is continuously distributed conditional on T, the control variable $U_M = F_{M|T}(M,T)$ recovers ε_M from observed data. Conditional on U_M, treatment exogeneity is restored: restriction (IV) ensures $T \perp\!\!\!\perp \varepsilon_Y \mid \varepsilon_M$.

However, conditioning on ε_M exacts a fatal cost for mediation analysis. The mediator $M = f_M(T, \varepsilon_M)$ becomes a deterministic function of T when ε_M is held fixed. In the linear model, $M = \gamma T + \varepsilon_M$ with ε_M constant reduces to $M = \gamma T + c$, making (T, M) perfectly collinear. Only the total effect $\tau + \theta\gamma$ is recoverable; the separate effects τ and θ are not. The failure is structural: identification of mediation parameters requires independent variation in T and M, and conditioning on ε_M eliminates the only source of such variation. After fixing ε_M, the mediator moves only because the treatment moves, and the data cannot distinguish a change in Y caused directly by T from one caused by the induced change in M.

Restriction (V): $\varepsilon_T \perp\!\!\!\perp \varepsilon_M \mid \varepsilon_Y$. Under this restriction, conditioning on ε_Y renders ε_T and ε_M independent. However, conditioning on the outcome error is fundamentally unworkable.

The control variable $U_Y = F_{Y|T,M}(Y,T,M)$ requires knowledge of the conditional distribution of Y given (T, M), which depends on the unknown structural parameters (τ, θ). Unlike $U_T = F_{T|Z}(T,Z)$, which is constructed from the first-stage equation whose parameter π_1 is identified by standard IV, U_Y cannot be constructed without already solving the identification problem. Moreover, conditioning on ε_Y makes $Y = f_Y(T, M, \varepsilon_Y)$ deterministic in (T, M), eliminating all stochastic variation in the out-

come. Finally, the pairs (t, m) observable at a given $U_Y = u$ satisfy the one-dimensional constraint $F_{Y|T,M}(y, t, m) = u$, collapsing the effective support from two dimensions to one. The partial derivatives $\partial f_Y/\partial t$ and $\partial f_Y/\partial m$ cannot be separately identified from variation along a single curve in (t, m) space.

C.4 Restriction (VI): A Valid Control Function Approach

The restriction $\varepsilon_M \perp\!\!\!\perp \varepsilon_Y \mid \varepsilon_T$ yields identification via the control function. The structural advantage is that ε_T enters the treatment equation—not the mediator or outcome equation—so conditioning on it does not destroy the variation needed for identification.

Identification. Suppose $f_T(z, \cdot)$ is strictly monotone in ε_T for each z and T is continuously distributed conditional on Z, so the control variable $U_T = F_{T|Z}(T, Z)$ recovers ε_T. Conditional on U_T:

(a) $T = f_T(Z, \varepsilon_T)$ becomes a deterministic function of Z, eliminating treatment endogeneity.

(b) $M = f_M(T, \varepsilon_M)$ retains independent variation through ε_M, which is not conditioned on. The regressors (T, M) are not collinear.

(c) Restriction (VI) ensures that, conditional on ε_T, the remaining errors ε_M and ε_Y are independent. Combined with (a), both T and M are exogenous:

$$(T, M) \perp\!\!\!\perp Y(t, m) \,\big|\, U_T \quad \text{for all } (t, m). \tag{100}$$

The causal effects are identified by:

$$\mathrm{E}\left[Y(t, m)\right] = \int_0^1 \mathrm{E}\left[Y \mid T = t,\, M = m,\, U_T = u\right] dF_{U_T}(u). \tag{101}$$

Linear model. In the linear mediation model, U_T is a monotone transformation of the first-stage residual $\hat{\varepsilon}_T = T - \hat{\pi}_1 Z$, and the augmented regression is:

$$Y = \tau\, T + \theta\, M + \rho\, \hat{\varepsilon}_T + \text{error}, \tag{102}$$

where ρ captures the endogeneity bias. OLS applied to (102) consistently estimates (τ, θ, ρ).

No 2SLS equivalent. In the standard single-endogenous-variable IV model ($Y = \beta T + \varepsilon_Y$), the control function estimator—OLS of Y on $(T, \hat{\varepsilon}_T)$—is numerically identical to 2SLS with instrument Z. This equivalence follows from the Frisch–Waugh–Lovell theorem: partialling out $\hat{\varepsilon}_T = T - \hat{\pi}_1 Z$ from T leaves exactly $\hat{T} = P_Z T$, the first-stage fitted values. This equivalence does *not* extend to the mediation model. In the augmented regression (102), partialling out $\hat{\varepsilon}_T$ yields

$$M_{\hat{\varepsilon}_T} Y = \tau\, M_{\hat{\varepsilon}_T} T + \theta\, M_{\hat{\varepsilon}_T} M + \text{error}, \tag{103}$$

where $M_{\hat{\varepsilon}_T} = I - \hat{\varepsilon}_T(\hat{\varepsilon}_T'\hat{\varepsilon}_T)^{-1}\hat{\varepsilon}_T'$ is the annihilator of $\hat{\varepsilon}_T$. Since $T = \hat{T} + \hat{\varepsilon}_T$ and $\hat{T} \perp \hat{\varepsilon}_T$ by construction, $M_{\hat{\varepsilon}_T} T = \hat{T}$: the treatment is replaced by its first-stage fitted values, exactly as in 2SLS.

However, $M_{\hat{\varepsilon}_T} M = M - (M'\hat{\varepsilon}_T/\hat{\varepsilon}_T'\hat{\varepsilon}_T)\hat{\varepsilon}_T \neq \hat{M}$: the mediator is not replaced by any instrumental variables projection. Instead, M is *purged of its correlation with the treatment error*—the component of mediator variation driven by the shared unobservable ε_T is removed, and the remaining variation (driven by ε_M, which is independent of ε_Y under restriction VI) is treated as exogenous.

The procedure is therefore a hybrid: T is instrumented via Z, while M is corrected by the control function rather than instrumented. There is no set of instruments that replicates (102) as a standard 2SLS regression of Y on (T, M), because the identification of θ does not rely on an instrument for M—it relies on the exogeneity of ε_M conditional on ε_T. Under MCA, by contrast, the collider mechanism (Section 3) generates a valid conditional instrument for M, and the mediation parameters are identified by a standard 2SLS system (Section 4.2). This distinction—instrumental variables versus control function correction—represents a fundamental difference in estimation paradigm between the two approaches.

Implementation. The procedure has five steps:

1. *First stage.* Regress T on (Z, X) to obtain fitted residuals $\hat{\varepsilon}_T = T - \hat{f}_T(Z, X)$.
2. *Augmented second stage.* Regress Y on $(T, M, \hat{\varepsilon}_T, X)$. The coefficients on T and M estimate τ and θ.
3. *Endogeneity test.* Test $H_0 : \rho = 0$ via the t-statistic on $\hat{\varepsilon}_T$. Rejection indicates the presence of treatment–outcome confounding that does not operate through the mediator.
4. *Generated-regressor correction.* Since $\hat{\varepsilon}_T$ is estimated, OLS standard errors from step 2 are inconsistent. Valid inference requires either an analytic variance correction or a bootstrap that resamples both stages jointly.
5. *Support verification.* Verify that (T, M) exhibits sufficient independent variation at each value of $\hat{\varepsilon}_T$. Limited overlap may lead to imprecise estimates.

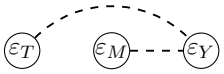
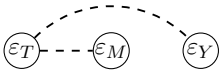
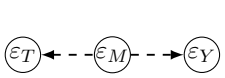
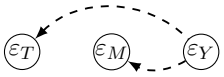
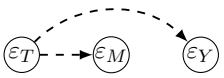
Additional requirements. Identification under restriction (VI) requires:

- *Monotonicity of f_T in ε_T*: For binary T and Z, this reduces to the Imbens and Angrist (1994) monotonicity condition $T(1) \geq T(0)$.
- *Continuity of T conditional on Z*: Required for the probability integral transform.
- *Support conditions*: For each $U_T = u$, the pair (t, m) must lie in the support of $(T, M) \mid U_T = u$.

None of these has an analogue under MCA.

Having analyzed both marginal and conditional restrictions, we can now summarize the full classification. Table C.1 summarizes the six candidate restrictions. The top row reports marginal independence restrictions, among which only mediated confounding (I) preserves the endogeneity structure and yields identification via the conditional IV argument. The bottom row reports conditional independence restrictions, among which only (VI) identifies the mediation parameters through the control-function approach. The remaining cases fail because they eliminate the variation required for identification or do not generate a valid instrument for the mediator.

Table C.1: Pairwise Independence Restrictions on the Structural Errors

(I) MCA: $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y$	**(II)**: $\varepsilon_T \perp\!\!\!\perp \varepsilon_M$	**(III)**: $\varepsilon_M \perp\!\!\!\perp \varepsilon_Y$
Identifies via conditional IV	**Fails: no T–M endogeneity**	**Fails: no M–Y endogeneity**
		
(IV): $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y \mid \varepsilon_M$	**(V)**: $\varepsilon_T \perp\!\!\!\perp \varepsilon_M \mid \varepsilon_Y$	**(VI)**: $\varepsilon_M \perp\!\!\!\perp \varepsilon_Y \mid \varepsilon_T$
Fails: no M variation	**Fails: no Y variation**	**Identifies via control function**
		

Notes: Each panel displays the dependence structure among $(\varepsilon_T, \varepsilon_M, \varepsilon_Y)$ under one restriction. Dashed lines indicate dependence; absent links indicate independence. Directed dashed arrows in the bottom row represent the conditional-independence structure. The observed causal structure $(Z \to T \to M \to Y,\ T \to Y)$ is the same in all six cases.

C.5 Why the Paper Adopts $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y$

Both MCA (restriction I: $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y$) and restriction (VI) ($\varepsilon_M \perp\!\!\!\perp \varepsilon_Y \mid \varepsilon_T$) identify the mediation parameters (τ, θ), but MCA is preferable on several grounds.

Interpretive transparency. MCA states that the unobserved determinants of treatment are independent of the unobserved determinants of the outcome—a condition directly interpretable in terms of the applied problem (Section 3.1). Restriction (VI) states that the mediator and outcome errors are independent *conditional on* the treatment error, requiring the researcher to reason about dependencies among three unobserved quantities simultaneously.

No auxiliary structural assumptions. MCA delivers conditional exogeneity and conditional relevance directly from the independence restriction, without requiring monotonicity of the first stage, continuity of the treatment, or support conditions (Proposition 1 and Section 3). In the linear model, MCA identifies population-level structural parameters via standard 2SLS even when T is binary. Restriction (VI) requires monotonicity, continuity, and support conditions to recover ε_T via the control function. When T is binary, exact inversion of the first stage fails, and identification requires parametric distributional assumptions or a propensity score reformulation.

Estimation simplicity. Under MCA in the linear model, the mediation parameters are identified by a just-identified GMM system estimated by standard 2SLS (Section 4.2). Under restriction (VI), estimation requires the two-step procedure (102) with a generated regressor, introducing additional estimation uncertainty.

Relative strength. Neither restriction implies the other. MCA restricts the marginal dependence between ε_T and ε_Y while leaving the conditional dependence $(\varepsilon_T, \varepsilon_Y) \mid \varepsilon_M$ unrestricted. Restriction (VI) leaves the marginal dependence unrestricted while restricting the conditional dependence. In the Gaussian model, MCA sets $\text{Corr}(\varepsilon_T, \varepsilon_Y) = 0$; restriction (VI) sets $\text{Corr}(\varepsilon_M, \varepsilon_Y \mid \varepsilon_T) = 0$ (the partial correlation of ε_M and ε_Y given ε_T). The plausibility of each depends on the application, but MCA's

marginal formulation is typically more natural in IV settings where the instrument is selected precisely because it affects the outcome only through the treatment.

D Proofs for Section 4

Proof of Theorem 1. Identification of π_1 and γ. Substituting (21) into the first moment of (31) gives $\sigma_{ZT} - \pi_1\sigma_{ZZ} = 0$, so $\pi_1 = \sigma_{ZT}/\sigma_{ZZ}$ under $\sigma_{ZZ} > 0$. Substituting (22) with $\mathrm{E}[Z\varepsilon_M] = 0$ into the second moment gives $\sigma_{ZM} - \gamma\sigma_{ZT} = 0$, so $\gamma = \sigma_{ZM}/\sigma_{ZT}$ under $\pi_1 \neq 0$ (which gives $\sigma_{ZT} = \pi_1\sigma_{ZZ} \neq 0$).

Identification of (τ, θ). The third moment of (31) expands to

$$\sigma_{ZY} - \tau\sigma_{ZT} - \theta\sigma_{ZM} = 0. \tag{104}$$

The fourth moment is $\mathrm{E}[(Y-\tau T-\theta M)(T-\pi_1 Z)] = 0$. At the true parameters, this equals $\mathrm{E}[\varepsilon_Y\varepsilon_T] = 0$ by MCA (17)—this is where MCA enters. Algebraically,

$$0 = \mathrm{E}\left[(Y - \tau T - \theta M)(T - \pi_1 Z)\right] = \left(\sigma_{TY} - \tau\sigma_{TT} - \theta\sigma_{TM}\right) - \underbrace{\left(\sigma_{ZY} - \tau\sigma_{ZT} - \theta\sigma_{ZM}\right)}_{=0 \text{ by } (104)}$$

which reduces to

$$\sigma_{TY} - \tau\sigma_{TT} - \theta\sigma_{TM} = 0. \tag{105}$$

Equations (104) and (105) form the linear system $\boldsymbol{A}(\tau, \theta)' = (\sigma_{ZY}, \sigma_{TY})'$ with

$$\boldsymbol{A} = \begin{pmatrix} \sigma_{ZT} & \sigma_{ZM} \\ \sigma_{TT} & \sigma_{TM} \end{pmatrix}.$$

Rank condition. Substituting the structural moments under $Z \perp\!\!\!\perp \varepsilon_M$ ($\sigma_{ZT} = \pi_1\sigma_{ZZ}$, $\sigma_{ZM} = \gamma\pi_1\sigma_{ZZ}$, $\sigma_{TM} = \gamma\sigma_{TT} + \mathrm{cov}(\varepsilon_T, \varepsilon_M)$), the determinant of $\boldsymbol{A}$ collapses cleanly:

$$\det(\boldsymbol{A}) = \sigma_{ZT}\,\sigma_{TM} - \sigma_{ZM}\,\sigma_{TT} = \pi_1\,\sigma_{ZZ} \cdot \mathrm{cov}(\varepsilon_T, \varepsilon_M),$$

where the $\gamma\,\pi_1\,\sigma_{ZZ}\,\sigma_{TT}$ contributions cancel. By the rank conditions (26), $\det(\boldsymbol{A}) \neq 0$. This is the algebraic content of the collider mechanism (Section 3): identification requires both an instrument with first-stage power and shared unobservables between T and M.

Closed forms. Cramer's rule applied to $\boldsymbol{A}(\tau, \theta)' = (\sigma_{ZY}, \sigma_{TY})'$ yields the unique solution (29). □

Proof of Corollary 1. Substituting (22) into (23) yields the reduced form (25): $Y = \tau^{\text{total}}T + \eta_Y$ with $\eta_Y = \theta\varepsilon_M + \varepsilon_Y$ and $\tau^{\text{total}} \equiv \tau + \theta\gamma$. By IV exogeneity (24), $\mathrm{E}[Z\eta_Y] = \theta\,\mathrm{E}[Z\varepsilon_M] + \mathrm{E}[Z\varepsilon_Y] = 0$. Multiplying the reduced form by Z and taking expectations gives $\sigma_{ZY} = \tau^{\text{total}}\,\sigma_{ZT}$, so $\tau^{\text{total}} = \sigma_{ZY}/\sigma_{ZT}$ under instrument relevance $\sigma_{ZT} \neq 0$. □

Remark. The argument does not invoke MCA: the total effect is identified by the standard IV reduced form regardless. MCA enters only in identifying the *decomposition* $\tau + \theta\gamma$ into separate direct and indirect components (Theorem 1). Corollary 1 therefore confirms internal consistency: the structural decomposition does not redefine the total effect, only its split.

Asymptotic Properties of the GMM Estimator

The following verification supports Remark 3. The result is a direct application of standard GMM asymptotic theory (Hayashi, 2000, Theorem 7.2); we verify the required conditions for the moment system (31).

Proof. Identification. By Theorem 1, the moment conditions (31) uniquely determine $\boldsymbol{\delta}_0$. The Jacobian $\boldsymbol{G} = \mathrm{E}[\partial \boldsymbol{g}_i(\boldsymbol{\delta})/\partial \boldsymbol{\delta}']\big|_{\boldsymbol{\delta}_0}$ has full column rank because the 4×4 system has a unique solution under instrument relevance ($\sigma_{ZT} \neq 0$) and the rank condition ($\pi_1 \, \sigma_{ZZ} \cdot \mathrm{Cov}(\varepsilon_{i,T}, \varepsilon_{i,M}) \neq 0$).

Asymptotic normality. Under i.i.d. sampling with $\mathrm{E}\,\|\boldsymbol{w}_i\|^4 < \infty$, the moment function $\boldsymbol{g}_i(\boldsymbol{w}_i|\boldsymbol{\delta})$ is square-integrable and $\boldsymbol{\Sigma}_{\boldsymbol{g}} = \mathrm{E}[\boldsymbol{g}_i \boldsymbol{g}_i']$ is the variance of the moment conditions. The Lindeberg–Lévy CLT yields $\sqrt{n}\,\bar{\boldsymbol{g}}_n(\boldsymbol{\delta}_0) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{g}})$, and the standard linearization argument gives $\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, (\boldsymbol{G}'\boldsymbol{\Sigma}_{\boldsymbol{g}}^{-1}\boldsymbol{G})^{-1})$.

Variance estimation. The sandwich estimator $\hat{\boldsymbol{\Sigma}}_n = n^{-1}(\boldsymbol{G}_n'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{g},n}^{-1}\boldsymbol{G}_n)^{-1}$ is consistent because $\boldsymbol{G}_n \xrightarrow{p} \boldsymbol{G}$ by the law of large numbers applied to $\partial \boldsymbol{g}_i/\partial \boldsymbol{\delta}'$, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{g},n} = n^{-1}\sum_{i=1}^n \boldsymbol{g}_i \boldsymbol{g}_i' \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{g}}$. Under conditional heteroskedasticity, the i.i.d. assumption can be replaced by stationarity, ergodicity, and a martingale difference condition on $\boldsymbol{g}_i$, with no change to the form of the estimator (Hayashi, 2000). □

E Treatment–Mediator Interaction

This appendix extends the linear mediation model of Section 4 by allowing the mediator effect to vary with treatment status. The baseline model (23) restricts the indirect effect to $\theta\gamma$, independent of t. We relax this restriction by augmenting the outcome equation with a treatment–mediator interaction.

E.1 The Interaction Model

Residualization convention. The interaction terms are formed in levels before residualization. In particular, form the raw products T_iM_i and Z_iT_i, residualize Y_i, T_i, M_i, Z_i, T_iM_i, and Z_iT_i with respect to $[\mathbf{1}, X]$, and suppress tildes afterward. Thus, throughout this appendix, T_iM_i denotes the residualized level interaction rather than the product of residualized T_i and residualized M_i; the same convention applies to Z_iT_i. This convention preserves the Frisch–Waugh–Lovell interpretation of the coefficients in the interacted model.

Under this convention, the residualized structural equations are:

$$T_i = \pi_1 \cdot Z_i + \varepsilon_{i,T}, \tag{106}$$

$$M_i = \gamma \cdot T_i + \varepsilon_{i,M}, \tag{107}$$

$$Y_i = \tau \cdot T_i + \theta \cdot M_i + \rho \cdot T_iM_i + \varepsilon_{i,Y}, \tag{108}$$

where ρ captures the treatment–mediator interaction: the marginal effect of M on Y at treatment level t is $\theta + \rho t$. The IV and MCA conditions are maintained:

$$Z_i \not\perp\!\!\!\perp T_i \quad \text{and} \quad Z_i \perp\!\!\!\perp (\varepsilon_{i,T}, \varepsilon_{i,M}, \varepsilon_{i,Y}), \qquad \varepsilon_{i,T} \perp\!\!\!\perp \varepsilon_{i,Y}. \tag{109}$$

The slope parameters of interest are $\boldsymbol{\delta}_0 = [\pi_1, \gamma, \tau, \theta, \rho]$.

E.2 Treatment-Dependent Indirect Effects

The underlying level model implies

$$M_i(1) - M_i(0) = \gamma, \qquad Y_i(t, m) - Y_i(t, m') = (\theta + \rho t)(m - m').$$

Hence the natural indirect effect at treatment level t is

$$\begin{aligned} \mathrm{NIE}(t) &= \mathrm{E}\left[Y_i(t, M_i(1)) - Y_i(t, M_i(0))\right] \\ &= (\theta + \rho t)\gamma. \end{aligned} \tag{110}$$

The interaction therefore separates the two indirect effects:

$$\mathrm{NIE}(0) = \theta\gamma, \tag{111}$$

$$\mathrm{NIE}(1) = (\theta + \rho)\gamma. \tag{112}$$

Their difference, $\mathrm{NIE}(1) - \mathrm{NIE}(0) = \rho\gamma$, measures how the mediator channel changes with treatment status. An aggregate total-effect decomposition also contains the interaction with the baseline level of

the mediator and is not needed for identifying the two indirect effects below.

E.3 Identification

Relative to the baseline linear outcome equation, the interaction adds the endogenous regressor $T_i M_i$. The additional instrument is $Z_i T_i$.

Proposition 8 (Orthogonality of $Z_i T_i$ and $\varepsilon_{i,Y}$)**.** *Under the IV conditions* (109) *and MCA,*

$$\mathrm{E}[Z_i T_i \cdot \varepsilon_{i,Y}] = 0.$$

Proof. Before residualization, the treatment equation makes T_i measurable with respect to $(Z_i, \varepsilon_{i,T})$. The IV conditions and MCA imply that $\varepsilon_{i,Y}$ is independent of $(Z_i, \varepsilon_{i,T})$. Hence $Z_i T_i \perp\!\!\!\perp \varepsilon_{i,Y}$ and $\mathrm{E}[Z_i T_i \varepsilon_{i,Y}] = 0$. Residualizing the level product $Z_i T_i$ on $[\mathbf{1}, X]$ preserves the corresponding partialled-out orthogonality condition. □

Adding this moment yields an exactly identified interacted outcome system.

Theorem 4 (Identification of the Interaction Model)**.** *Suppose the IV conditions* (109)*, MCA, instrument variation, and instrument relevance hold. If the matrix* $\boldsymbol{A}$ *in* (114) *is nonsingular, then the five slope parameters* $\boldsymbol{\delta}_0 = [\pi_1, \gamma, \tau, \theta, \rho]$ *are identified by*

$$\mathrm{E}[\boldsymbol{g}_i(\boldsymbol{w}_i \mid \boldsymbol{\delta}_0)] = \mathbf{0}, \qquad \boldsymbol{g}_i \equiv \begin{pmatrix} Z_i \cdot (T_i - \pi_1 Z_i) \\ Z_i \cdot (M_i - \gamma T_i) \\ Z_i \cdot (Y_i - \tau T_i - \theta M_i - \rho T_i M_i) \\ T_i \cdot (Y_i - \tau T_i - \theta M_i - \rho T_i M_i) \\ Z_i T_i \cdot (Y_i - \tau T_i - \theta M_i - \rho T_i M_i) \end{pmatrix}. \tag{113}$$

The first two moments identify

$$\pi_1 = \frac{\sigma_{ZT}}{\sigma_{ZZ}}, \qquad \gamma = \frac{\sigma_{ZM}}{\sigma_{ZT}},$$

where $\sigma_{AB} \equiv \mathrm{E}[A_i B_i]$ *for the residualized variables. The remaining parameters* (τ, θ, ρ) *solve*

$$\underbrace{\begin{pmatrix} \sigma_{ZT} & \sigma_{ZM} & \sigma_{Z,TM} \\ \sigma_{TT} & \sigma_{TM} & \sigma_{T^2M} \\ \sigma_{ZT^2} & \sigma_{ZTM} & \sigma_{ZT^2M} \end{pmatrix}}_{\boldsymbol{A}} \begin{pmatrix} \tau \\ \theta \\ \rho \end{pmatrix} = \begin{pmatrix} \sigma_{ZY} \\ \sigma_{TY} \\ \sigma_{ZTY} \end{pmatrix}, \tag{114}$$

provided $\det(\boldsymbol{A}) \neq 0$*. The mixed-moment shorthand in* (114) *follows the residualization convention above: for example,* σ_{ZT^2} *is the cross moment between the residualized level product* $Z_i T_i$ *and residualized* T_i*.*

Proof. Validity. The IV conditions give the first three moments in (113). MCA implies

$$\mathrm{E}[T_i \varepsilon_{i,Y}] = \pi_1 \mathrm{E}[Z_i \varepsilon_{i,Y}] + \mathrm{E}[\varepsilon_{i,T} \varepsilon_{i,Y}] = 0,$$

which gives the fourth moment. Proposition 8 gives the fifth moment.

Identification. The first two moments imply $\sigma_{ZT} = \pi_1 \sigma_{ZZ}$ and $\sigma_{ZM} = \gamma \sigma_{ZT}$, identifying (π_1, γ) under instrument variation and relevance. Substituting (108) into the last three moments yields (114). Nonsingularity of $\boldsymbol{A}$ then identifies (τ, θ, ρ). □

The primitive rank condition is $\det(\boldsymbol{A}) \neq 0$. Equivalently, after partialling out the included exogenous regressors $(1, T_i, X_i)$ in the level interacted regression, the excluded instruments $(Z_i, Z_i T_i)$ must have full rank for the endogenous regressors $(M_i, T_i M_i)$. This condition should be checked through the interacted first stages in estimation.

E.4 Estimation

The outcome equation can be estimated by 2SLS with T_i included exogenously, $(M_i, T_i M_i)$ treated as endogenous, and $(Z_i, Z_i T_i)$ used as excluded instruments. In the residualized notation of this appendix, the first stages and second stage are

$$\text{First stage for } M_i: \quad M_i = \delta_1 T_i + \delta_2 Z_i + \delta_3 Z_i T_i + \eta_{i,M}, \tag{115}$$

$$\text{First stage for } T_i M_i: \quad T_i M_i = \xi_1 T_i + \xi_2 Z_i + \xi_3 Z_i T_i + \eta_{i,TM}, \tag{116}$$

$$\text{Second stage:} \quad Y_i = \tau T_i + \theta \widehat{M}_i + \rho \widehat{T_i M_i} + \varepsilon_{i,Y}. \tag{117}$$

In implementation, $T_i M_i$ and $Z_i T_i$ are formed in levels before the intercept and covariates are partialled out. Moreover, $\widehat{T_i M_i}$ in (117) is the fitted value from the first stage for $T_i M_i$, not the product $T_i \widehat{M}_i$. The first-stage diagnostics for both endogenous regressors should be reported.

F Sensitivity Analysis: Formulas and Implementation

This appendix provides proofs, formulas, and implementation guidelines for the sensitivity frameworks. We first prove Proposition 3 and describe the estimation algorithm for the κ-based analysis developed in Section 5.2. We then present the complementary R^2-based sensitivity framework adapted from Cinelli and Hazlett (2024), including its definitions, bias and standard error factors, Anderson–Rubin integration, and robustness values. We conclude with step-by-step implementation guidelines for both frameworks and benchmark bounds for κ.

κ-Based Sensitivity: Proof and Estimation

Proof of Proposition 3. Maintain the linear model (21)–(23) with instrument exogeneity and the rank condition $\text{Cov}(\varepsilon_T, \varepsilon_M) \neq 0$, but allow $\kappa \equiv \text{Cov}(\varepsilon_T, \varepsilon_Y) \neq 0$.

Part (a). Working in the residualized variables of (21)–(23), X has already been partialled out; let tildes denote residuals from the additional projection on $[\mathbf{1}, T]$. Throughout, write $\sigma_Z^2 \equiv \sigma_{ZZ}$, $\sigma_{\varepsilon_T}^2 \equiv \text{E}[\varepsilon_T^2]$, and $\rho_{TM} \equiv \text{Cov}(\varepsilon_T, \varepsilon_M)$ (a covariance, not a correlation). The probability limit of $\hat{\theta}$ is $\text{Cov}(\tilde{Z}, \tilde{Y})/\text{Cov}(\tilde{Z}, \tilde{M})$. Substituting $Y = \tau T + \theta M + \varepsilon_Y$,

$$\text{Cov}(\tilde{Z}, \tilde{Y}) = \theta \cdot \text{Cov}(\tilde{Z}, \tilde{M}) + \text{Cov}(\tilde{Z}, \varepsilon_Y).$$

The FWL projection coefficient is $\beta_{TZ} \equiv \text{Cov}(Z, T)/\text{Var}(T) = \pi_1 \sigma_Z^2/(\pi_1^2 \sigma_Z^2 + \sigma_{\varepsilon_T}^2)$, using $T = \pi_1 Z + \varepsilon_T$. Since $\text{Cov}(Z, \varepsilon_Y) = 0$ and $\text{Cov}(T, \varepsilon_Y) = \text{Cov}(\varepsilon_T, \varepsilon_Y) = \kappa$,

$$\text{Cov}(\tilde{Z}, \varepsilon_Y) = \text{Cov}(Z, \varepsilon_Y) - \beta_{TZ}\text{Cov}(T, \varepsilon_Y) = -\frac{\pi_1 \sigma_Z^2}{\pi_1^2 \sigma_Z^2 + \sigma_{\varepsilon_T}^2} \cdot \kappa.$$

An analogous calculation, using $M = \gamma T + \varepsilon_M$ and $\text{Cov}(T, \varepsilon_M) = \rho_{TM}$, gives

$$\text{Cov}(\tilde{Z}, \tilde{M}) = -\frac{\pi_1 \sigma_Z^2}{\pi_1^2 \sigma_Z^2 + \sigma_{\varepsilon_T}^2} \cdot \rho_{TM}.$$

The common factor $-\pi_1 \sigma_Z^2/(\pi_1^2 \sigma_Z^2 + \sigma_{\varepsilon_T}^2)$ cancels in the ratio (it is nonzero by $\sigma_{ZZ} > 0$ and $\pi_1 \neq 0$):

$$\text{plim}(\hat{\theta}) - \theta = \frac{\text{Cov}(\tilde{Z}, \varepsilon_Y)}{\text{Cov}(\tilde{Z}, \tilde{M})} = \frac{\kappa}{\rho_{TM}},$$

establishing (42).

Part (b). The plug-in estimator uses $\hat{\varepsilon}_T = T - \hat{\pi}_1 Z$ and $\hat{\varepsilon}_M = M - \hat{\gamma}_{\text{IV}} T$, where $\hat{\gamma}_{\text{IV}} = \widehat{\text{Cov}}(Z, M)/\widehat{\text{Cov}}(Z, T)$. Consistency of $\hat{\pi}_1$ for π_1 follows from $\text{Cov}(Z, \varepsilon_T) = 0$, which is a maintained condition unaffected by κ. For $\hat{\gamma}_{\text{IV}}$: the numerator $\widehat{\text{Cov}}(Z, M) \xrightarrow{p} \text{Cov}(Z, M) = \gamma \cdot \text{Cov}(Z, T) + \text{Cov}(Z, \varepsilon_M) = \gamma \pi_1 \sigma_Z^2$, since $\text{Cov}(Z, \varepsilon_M) = 0$ is maintained regardless of κ. Similarly $\widehat{\text{Cov}}(Z, T) \xrightarrow{p} \pi_1 \sigma_Z^2$, so $\hat{\gamma}_{\text{IV}} \xrightarrow{p} \gamma$. Therefore $\hat{\varepsilon}_M \xrightarrow{p} M - \gamma T = \varepsilon_M$, and $\widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M) \xrightarrow{p} \text{Cov}(\varepsilon_T, \varepsilon_M) = \rho_{TM}$ by the continuous mapping theorem.

Part (c). By (b), $\widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M) \xrightarrow{p} \rho_{TM}$. By part (a), $\hat{\theta} \xrightarrow{p} \theta + \kappa/\rho_{TM}$. The breakeven $\hat{\kappa}^* = \hat{\theta} \cdot \widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M) \xrightarrow{p} (\theta + \kappa/\rho_{TM}) \cdot \rho_{TM} = \theta \rho_{TM} + \kappa = \kappa^* + \kappa$, where $\kappa^* \equiv \theta \rho_{TM}$. Under MCA ($\kappa = 0$), $\hat{\kappa}^* \xrightarrow{p} \kappa^*$. Since $\hat{\kappa}^*$ is a smooth function of $(\hat{\theta}, \widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M))$, both of which are $\sqrt{n}$-asymptotically normal, the delta method gives $\sqrt{n}(\hat{\kappa}^* - \kappa^*) \xrightarrow{d} N(0, V_{\kappa^*})$, where $V_{\kappa^*} = \nabla' \Sigma_{\theta,\rho} \nabla$, with gradient

$\nabla = (\rho_{TM},\, \theta)'$ and $\Sigma_{\theta,\rho}$ the joint asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta,\, \widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M) - \rho_{TM})'$. In the general case ($\kappa \neq 0$), $\hat{\kappa}^*$ is biased upward by κ, reflecting the fact that the biased estimate $\hat{\theta}$ attributes both the true effect and the bias to the breakeven threshold.

Part (d). When $\kappa_0 = \kappa$, the adjusted estimate $\hat{\theta}_{\text{adj}}(\kappa_0) = \hat{\theta} - \kappa_0/\widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$ has probability limit $\theta + \kappa/\rho_{TM} - \kappa/\rho_{TM} = \theta$, establishing $\sqrt{n}$-consistency. For asymptotic normality, expand:

$$\sqrt{n}\big(\hat{\theta}_{\text{adj}}(\kappa_0) - \theta\big) = \sqrt{n}\big(\hat{\theta} - \text{plim}(\hat{\theta})\big) + \kappa \cdot \frac{\sqrt{n}\big(\widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M) - \rho_{TM}\big)}{\rho_{TM}\, \widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)}.$$

Both terms are $O_p(1)$ and jointly asymptotically normal. When $\kappa = 0$, the second term vanishes and $\sqrt{n}(\hat{\theta}_{\text{adj}} - \theta)/\sqrt{V_{\hat{\theta}}} \xrightarrow{d} N(0,1)$, so $\text{CI}_{\kappa_0}(\theta)$ with $\widehat{\text{SE}}(\hat{\theta})$ achieves coverage $1-\alpha$. When $\kappa \neq 0$, the second term contributes additional asymptotic variance $\kappa^2 V_{\hat{\rho}}/\rho_{TM}^4$ (plus cross-terms), so $\widehat{\text{SE}}(\hat{\theta})$ understates the true variability and the interval (46) may undercover. Exact coverage for $\kappa \neq 0$ requires the delta-method standard error of $h(\hat{\theta}, \widehat{\text{Cov}}) = \hat{\theta} - \kappa_0/\widehat{\text{Cov}}$, with gradient $(1,\ \kappa_0/\widehat{\text{Cov}}^2)'$ evaluated at $(\text{plim}(\hat{\theta}),\, \rho_{TM})$. □

Estimation Algorithm. The κ-based sensitivity analysis requires estimating $\text{Cov}(\varepsilon_T, \varepsilon_M)$. The following algorithm is valid regardless of whether MCA holds:

1. Estimate the first stage by OLS: $T_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i + \hat{\pi}_X' X_i + \hat{\varepsilon}_{i,T}$.
2. Estimate γ by IV, using Z as instrument for T in $M_i = \gamma T_i + \gamma_X' X_i + \varepsilon_{i,M}$. Equivalently, compute $\hat{\gamma}_{\text{IV}} = \widehat{\text{Cov}}(Z, M \mid X)/\widehat{\text{Cov}}(Z, T \mid X)$ and set $\hat{\varepsilon}_{i,M} = M_i - \hat{\gamma}_{\text{IV}} T_i - \hat{\gamma}_X' X_i$.
3. Compute $\widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M) = n^{-1} \sum_{i=1}^n \hat{\varepsilon}_{i,T} \hat{\varepsilon}_{i,M}$.
4. Compute the signed breakeven $\hat{\kappa}^* = \hat{\theta} \cdot \widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$ and the signed inference breakeven $\hat{\kappa}^*_{\text{CI}}$ from (43)–(44); report both the magnitude $|\hat{\kappa}^*|$ and the sign $\text{sign}(\hat{\kappa}^*)$.
5. For a hypothesized violation κ_0, compute $\hat{\theta}_{\text{adj}}(\kappa_0)$ and $\text{CI}_{\kappa_0}(\theta)$ from (45)–(46).

The key identification result for Step 2 is that $\text{Cov}(Z, \varepsilon_M) = 0$ holds under the maintained instrument exogeneity regardless of whether $\kappa = 0$, so the IV estimator of γ is consistent even when MCA fails. This ensures that $\widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$ is consistent for ρ_{TM} uniformly over κ (Proposition 3(b)).

R^2-Based Sensitivity for the Conditional Reduced Form

The κ-based framework targets MCA violations directly. A complementary threat is that an omitted confounder W affects both the residualized instrument $\tilde{Z}$ and the residualized outcome $\tilde{Y}$ in data orthogonalized with respect to $[\mathbf{1}, X, T]$. This section adapts the OVB framework of Cinelli and Hazlett (2024) to assess robustness of the conditional reduced form to such confounders.

Importantly, this framework does *not* directly test $\kappa = 0$. MCA violations operate through the treatment equation: after partialling out T, ε_T retains substantial partial R^2 with $\tilde{Z}$ but essentially zero partial R^2 with $\tilde{Y}$ conditional on $\tilde{Z}$, because the channel $\varepsilon_T \to T \to Y$ is absorbed by conditioning on T. The R^2-based robustness values are therefore informative about confounders of the reduced form (e.g., violations of instrument exogeneity operating through channels other than T), but uninformative

about MCA itself. The κ-based analysis targets MCA violations directly; the R^2-based analysis targets a complementary threat. Section 7, Experiment 3 provides simulation evidence confirming this distinction.

The residualized estimator $\hat{\theta}$ from (34) equals the ratio of two OLS coefficients:

$$\text{Conditional reduced form:} \quad \tilde{Y} = \hat{\lambda}_{\text{res}} \tilde{Z} + \hat{\varepsilon}_{y,\text{res}}, \tag{118}$$

$$\text{Conditional first stage:} \quad \tilde{M} = \hat{\phi}_{\text{res}} \tilde{Z} + \hat{\varepsilon}_{m,\text{res}}, \tag{119}$$

so that $\hat{\theta} = \hat{\lambda}_{\text{res}} / \hat{\phi}_{\text{res}}$. Since each component is a standard OLS regression, the OVB framework of Cinelli and Hazlett (2024) applies to each individually. The strength of a hypothetical confounder W is parameterized by two partial R^2 measures:[25]

$$\begin{aligned} R^2_{\tilde{Y} \sim W | \tilde{Z}} &\equiv \text{partial } R^2 \text{ of } W \text{ with } \tilde{Y} \text{ after controlling for } \tilde{Z}, \\ R^2_{\tilde{Z} \sim W} &\equiv \text{partial } R^2 \text{ of } W \text{ with } \tilde{Z}. \end{aligned} \tag{120}$$

In applications, the strength of potential confounders can be bounded by comparison to observed covariates X_j via the relative strength parameters k_Z and k_Y:

$$k_Z \equiv \frac{R^2_{Z \sim W | X_{-j}, T}}{R^2_{Z \sim X_j | X_{-j}, T}}, \qquad k_Y \equiv \frac{R^2_{Y \sim W | Z, X_{-j}, T}}{R^2_{Y \sim X_j | Z, X_{-j}, T}}. \tag{121}$$

Setting $k_Z = k_Y = 1$ corresponds to a confounder as strong as X_j.

Robustness Values for the Mediator Effect. Building on Cinelli and Hazlett (2024), we define summary measures of the R^2-based sensitivity. Let $q^* \in [0, 1]$ denote the hypothesized percent reduction in the point estimate due to confounding; $q^* = 1$ corresponds to reducing $\hat{\theta}$ to zero.

Definition 2 (Extreme Robustness Value)**.** *The* extreme robustness value $XRV_{q^*,\alpha}(\hat{\theta})$ *is the minimum partial* R^2 *that the omitted variable* W *must have with the instrument* $\tilde{Z}$ *alone in order to change the conclusion at the* α *level:*

$$XRV_{q^*,\alpha}(\hat{\theta}) \equiv \inf \left\{ R^2_{\tilde{Z} \sim W} : \ (1 - q^*)\hat{\theta} \in CI^{\max}_{1-\alpha, 1, R^2_{\tilde{Z} \sim W}}(\theta) \right\}, \tag{122}$$

where the subscript 1 *indicates that* $R^2_{\tilde{Y} \sim W | \tilde{Z}}$ *is unconstrained (i.e., set to its upper bound of* 1*).*

Definition 3 (Robustness Value)**.** *The* robustness value $RV_{q^*,\alpha}(\hat{\theta})$ *is the minimum partial* R^2 *that* W *must have with* both $\tilde{Z}$ *and* $\tilde{Y}$ *simultaneously to change the conclusion:*

$$RV_{q^*,\alpha}(\hat{\theta}) \equiv \inf \left\{ RV : \ (1 - q^*)\hat{\theta} \in CI^{\max}_{1-\alpha, RV, RV}(\theta) \right\}. \tag{123}$$

Both statistics have direct interpretations. An $\text{RV}_{1,0.05}(\hat{\theta}) = 0.03$ means that a confounder explaining at least 3% of the residual variation in both $\tilde{Z}$ and $\tilde{Y}$ (conditional on T, X) is needed to bring the 95% confidence interval for θ to include zero.

[25] All partial R^2 measures are computed in the residualized data after projecting out $[\mathbf{1}, X, T]$. The notation $R^2_{\tilde{Y} \sim W | \tilde{Z}}$ corresponds to $R^2_{Y \sim W | Z, X, T}$ in the notation of Cinelli and Hazlett (2024).

A key decomposition follows from the ratio structure $\hat{\theta} = \hat{\lambda}_{\text{res}}/\hat{\phi}_{\text{res}}$. Because the AR confidence set for θ is the intersection of the half-planes determined by the significance of $\hat{\phi}_{\theta^*}$ (the AR statistic at the hypothesized value) and $\hat{\phi}_{\text{res}}$ (the conditional first stage), confounding can overturn $\hat{\theta}$ if and only if it overturns at least one of these two components:

$$\text{XRV}_{\geq q^*,\alpha}(\hat{\theta}) = \min\left\{\text{XRV}_{1,\alpha}(\hat{\phi}_{\theta^*}),\ \ \text{XRV}_{1,\alpha}(\hat{\phi}_{\text{res}})\right\}, \tag{124}$$

$$\text{RV}_{\geq q^*,\alpha}(\hat{\theta}) = \min\left\{\text{RV}_{1,\alpha}(\hat{\phi}_{\theta^*}),\ \ \text{RV}_{1,\alpha}(\hat{\phi}_{\text{res}})\right\}, \tag{125}$$

where $\theta^* \equiv (1 - q^*)\hat{\theta}$ is the hypothesized true value under q^*-percent confounding. The conditional first stage, powered by the collider mechanism (Section 3.3), is typically weaker than a direct instrument channel, so the first-stage component often binds.

Bias and Standard Error Factors. Given postulated values of $(R^2_{\tilde{Y}\sim W|\tilde{Z}},\, R^2_{\tilde{Z}\sim W})$, the bias in the conditional reduced form coefficient $\hat{\lambda}_{\text{res}}$ and the adjustment to its standard error are characterized by Theorem 1 of Cinelli and Hazlett (2024). Let $\tilde{Y}^{\perp\tilde{Z}}$ denote the OLS residual from regressing $\tilde{Y}$ on $\tilde{Z}$. Define the bias factor and standard error factor:

$$\text{BF} \equiv \sqrt{R^2_{\tilde{Y}\sim W|\tilde{Z}} \cdot \frac{R^2_{\tilde{Z}\sim W}}{1 - R^2_{\tilde{Z}\sim W}}} \times \frac{\text{sd}(\tilde{Y}^{\perp\tilde{Z}})}{\text{sd}(\tilde{Z})}, \tag{126}$$

$$\text{SEF} \equiv \sqrt{\frac{1 - R^2_{\tilde{Y}\sim W|\tilde{Z}}}{1 - R^2_{\tilde{Z}\sim W}}} \times \frac{\text{sd}(\tilde{Y}^{\perp\tilde{Z}})}{\text{sd}(\tilde{Z})} \times \frac{1}{\sqrt{\text{df} - 1}}, \tag{127}$$

where $\text{df} = n - K_{\text{controls}} - 1$ and K_{controls} counts the conditioning variables (including T and X). The bias-adjusted critical value is:

$$t^{\dagger}_{\alpha,\text{df}-1,\boldsymbol{R}^2} \equiv \text{SEF}\sqrt{\frac{\text{df}}{\text{df} - 1}} \times t^{*}_{\alpha,\text{df}-1} + \text{BF}\sqrt{\text{df}}, \tag{128}$$

where $\boldsymbol{R}^2 = (R^2_{\tilde{Y}\sim W|\tilde{Z}},\, R^2_{\tilde{Z}\sim W})$ and $t^{*}_{\alpha,\text{df}-1}$ is the standard critical value. When $\boldsymbol{R}^2 = (0, 0)$, the adjusted value reduces to the standard one.

These adjusted critical values integrate with the Anderson–Rubin (AR) framework (Anderson and Rubin, 1949) to produce bias-adjusted confidence intervals. Under $H_0 : \theta = \theta_0$, construct $Y_{\theta_0} \equiv \tilde{Y} - \theta_0\tilde{M}$ and regress on $\tilde{Z}$; let $\hat{\phi}_{\theta_0}$ denote the resulting OLS coefficient and $t_{\hat{\phi}_{\theta_0}}$ its t-statistic. The bias-adjusted confidence interval for the mediator effect is:

$$\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\theta) \equiv \left\{\theta_0 :\ t^2_{\hat{\phi}_{\theta_0}} \leq \left(t^{\dagger\ \max}_{\alpha,\text{df}-1,\boldsymbol{R}^2}\right)^2\right\}, \tag{129}$$

where $t^{\dagger\ \max}_{\alpha,\text{df}-1,\boldsymbol{R}^2}$ denotes the maximum of $t^{\dagger}_{\alpha,\text{df}-1,\boldsymbol{R}^2}$ over all omitted variables whose partial R^2 with $\tilde{Y}$ (given $\tilde{Z}$) is bounded by $R^2_{\tilde{Y}\sim W|\tilde{Z}}$ and whose partial R^2 with $\tilde{Z}$ is bounded by $R^2_{\tilde{Z}\sim W}$. The maximum is attained when the bias and variance components reinforce each other; see Theorem 3 of Cinelli and Hazlett (2024).

Robustness Values for Individual OLS Coefficients

The XRV and RV for $\hat{\theta}$ are computed via the decomposition (124)–(125) in the main text, which reduces the problem to computing XRV and RV for individual OLS coefficients (the conditional reduced form coefficient $\hat{\lambda}_{\text{res}}$ and the conditional first stage coefficient $\hat{\phi}_{\text{res}}$). For any OLS coefficient $\hat{\beta}$ with t-statistic $t_{\hat{\beta}}$ and f-statistic $f_{\hat{\beta}} \equiv t_{\hat{\beta}}^2$, these are given in closed form by Theorems 4–5 and 7–8 of Cinelli and Hazlett (2024). Define $f_{q^*,\alpha}(\hat{\beta}) \equiv q^*|f_{\hat{\beta}}| - f^*_{\alpha,\text{df}-1}$, where $f^*_{\alpha,\text{df}-1} \equiv (t^*_{\alpha,\text{df}-1})^2/(\text{df}-1)$. The extreme robustness value is:

$$\text{XRV}_{1,\alpha}(\hat{\beta}) \equiv f_{1,\alpha}(\hat{\beta}) = |f_{\hat{\beta}}| - f^*_{\alpha,\text{df}-1}, \tag{130}$$

which is simply the excess f-statistic above the critical threshold, expressed as a partial R^2.[26] The robustness value is:

$$\text{RV}_{1,\alpha}(\hat{\beta}) \equiv \begin{cases} \frac{1}{2}\left(\sqrt{f^4_{1,\alpha}(\hat{\beta}) + 4f^2_{1,\alpha}(\hat{\beta})} - f^2_{1,\alpha}(\hat{\beta})\right), & \text{if } f^*_{\alpha,\text{df}-1} < f_{1,\alpha}(\hat{\beta}) < (f^*_{\alpha,\text{df}-1})^{-1}, \\ \text{XRV}_{1,\alpha}(\hat{\beta}), & \text{otherwise.} \end{cases} \tag{131}$$

The first case is an interior solution where the constraint that $R^2_{\tilde{Y}\sim W|\tilde{Z}} = R^2_{\tilde{Z}\sim W}$ binds; the second case arises when the constraint is not binding and the RV reduces to the XRV. These closed-form expressions are applied to $\hat{\phi}_{\theta^*}$ and $\hat{\phi}_{\text{res}}$ in (124)–(125).[27]

Practical Implementation

The sensitivity analysis proceeds in five steps, organized by framework.[28]

1. **Estimate the mediation model.** Run the 2SLS regressions of Section 4.2. Extract: the mediator effect estimate $\hat{\theta}$ with standard error $\widehat{\text{SE}}(\hat{\theta})$, the conditional first stage (32) with its F-statistic, and the conditional reduced form (118).

2. **κ-based sensitivity.** Compute the breakeven values and adjusted inference:

   (a) Estimate $\widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$ using the algorithm above: OLS residuals from the first stage, IV residuals from the mediator equation. Under non-uniform observation weights w_i (analytic or sampling weights), use the Hajek (self-normalizing) form

$$\widehat{\text{Cov}}_w(\hat{\varepsilon}_T, \hat{\varepsilon}_M) = \frac{1}{\sum_i w_i} \sum_{i=1}^{n} w_i \, \hat{\varepsilon}_{i,T} \, \hat{\varepsilon}_{i,M}, \tag{132}$$

[26] When $\text{XRV}_{1,\alpha}(\hat{\beta}) \leq 0$, the estimate is not significant even without confounding.

[27] The closed-form RV/XRV expressions of Cinelli and Hazlett (2024) are derived under homoskedastic OLS. When inference uses heteroskedasticity-robust, cluster-robust, or spatial standard errors, the corresponding f-statistic $f_{\hat{\beta}} = t^2_{\hat{\beta}}$ is computed with the SE-regime-appropriate variance estimator, following the convention adopted by the `sensemakr` (Cinelli and Hazlett, 2020) and `iv.sensemakr` R packages. This substitution preserves the interpretation of the RV as a partial-R^2 threshold to leading order but is a documented theoretical approximation rather than an exact result; researchers should report the SE regime alongside the RV.

[28] The R^2-based computations can be performed using the R package `sensemakr` (Cinelli and Hazlett, 2020) for standard OVB sensitivity, and its IV extension `iv.sensemakr` for the Anderson–Rubin-based analysis. The κ-based computations require only standard IV estimation and residual covariance calculation.

which is consistent for the population covariance under non-uniform weights and reduces to $n^{-1}\sum_i \hat{\varepsilon}_{i,T}\hat{\varepsilon}_{i,M}$ when $w_i \equiv 1$. The same substitution applies to all subsequent residual covariances and standard deviations in this appendix; in particular, $\hat{\sigma}^2_{\varepsilon_T}$, $\hat{\sigma}^2_{\varepsilon_Y}$, and the benchmark covariances $\hat{\kappa}_j = \widehat{\mathrm{Cov}}(\hat{\varepsilon}_T, X_j)$ are computed in Hajek form, and the residuals are obtained from weighted least squares versions of the first-stage and mediator equations.

(b) Report the signed breakeven $\hat{\kappa}^* = \hat{\theta} \cdot \widehat{\mathrm{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$ and the signed inference breakeven $\hat{\kappa}^*_{\mathrm{CI}} = (\hat{\theta} - z_{\alpha/2} \cdot \widehat{\mathrm{SE}}(\hat{\theta})) \cdot \widehat{\mathrm{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$, reporting magnitude and sign separately.

(c) Report the signed breakeven as a correlation: $\hat{\kappa}^*/(\hat{\sigma}_{\varepsilon_T}\hat{\sigma}_{\varepsilon_Y})$, which is bounded in $[-1, 1]$.

(d) For hypothesized violations κ_0 (e.g., calibrated from observed covariances $\widehat{\mathrm{Cov}}(\hat{\varepsilon}_T, X_j)$), report the adjusted estimate $\hat{\theta}_{\mathrm{adj}}(\kappa_0)$ and confidence interval $\mathrm{CI}_{\kappa_0}(\theta)$ from (45)–(46).

(e) Construct the *κ-adjusted reporting table* by reporting $\hat{\theta}_{\mathrm{adj}}(\kappa_0)$ and $\mathrm{CI}_{\kappa_0}(\theta)$ at a grid of fractions $\kappa_0 = f \cdot \hat{\kappa}^*$ for $f \in \{0, 0.10, 0.25, 0.50, 0.75, 0.90, 1.00\}$. The point estimate satisfies the closed-form linear interpolation

$$\hat{\theta}_{\mathrm{adj}}(f \cdot \hat{\kappa}^*) = \hat{\theta} \cdot (1 - f), \tag{133}$$

which follows immediately from (45) and the definition of $\hat{\kappa}^*$. The boundary values reproduce the baseline estimate ($f = 0 \Rightarrow \hat{\theta}_{\mathrm{adj}} = \hat{\theta}$) and zero ($f = 1 \Rightarrow \hat{\theta}_{\mathrm{adj}} = 0$); intermediate rows trace how the conclusion moves as MCA is partially relaxed. This table is the recommended primary user-facing display of κ-sensitivity.

(f) Compare $\hat{\kappa}^*$ to the LR estimate $\hat{\kappa}$ when $K \geq 2$ (Section 5.1).

Interpretation: Large $|\hat{\kappa}^*|$ relative to plausible values of κ indicates robustness of the mediator effect to MCA violations. A low conditional F-statistic (Remark 2) signals that $|\hat{\kappa}^*|$ will be small, indicating high sensitivity.

Warning and fallback rule: When $\widehat{\mathrm{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M) \approx 0$ (equivalently, the conditional first-stage F-statistic falls below the Olea–Pflueger (Montiel Olea and Pflueger, 2013) critical value $c(K, \tau, \alpha)$ at the chosen tolerance τ), the κ-adjusted estimate $\hat{\theta}_{\mathrm{adj}}(\kappa_0)$ involves division by a near-zero quantity and is unreliable. Implementations should fall back to reporting the baseline estimate $\hat{\theta}$ alone in the κ-adjusted reporting table, flag the fallback explicitly, and direct inference to the Anderson–Rubin confidence set from Section 5.1 or the R^2-based bounds (Step 3). The breakeven correlation $\hat{\rho}^*_{TY}$ remains informative when $|\hat{\rho}^*_{TY}| > 1$, since it certifies that no feasible error correlation can overturn the result regardless of conditional first-stage strength.

3. **R^2-based robustness values.** Using the decomposition in equations (124)–(125), compute:

$$\mathrm{XRV}_{\geq q^*,\alpha}(\hat{\theta}) = \min\left\{\mathrm{XRV}_{1,\alpha}(\hat{\phi}_{\theta^*}),\ \mathrm{XRV}_{1,\alpha}(\hat{\phi}_{\mathrm{res}})\right\},$$
$$\mathrm{RV}_{\geq q^*,\alpha}(\hat{\theta}) = \min\left\{\mathrm{RV}_{1,\alpha}(\hat{\phi}_{\theta^*}),\ \mathrm{RV}_{1,\alpha}(\hat{\phi}_{\mathrm{res}})\right\},$$

where $\hat{\phi}_{\theta^*}$ is the OLS coefficient from regressing $\tilde{Y} - \theta^*\tilde{M}$ on $\tilde{Z}$ with $\theta^* = (1 - q^*)\hat{\theta}$, and $\hat{\phi}_{\mathrm{res}}$ is the conditional first-stage coefficient from (119). Report a sensitivity table with the following columns for each of the two component regressions (AR statistic, conditional first stage): point

estimate, t-value, $\text{XRV}_{1,\alpha}$, and $\text{RV}_{1,\alpha}$. Report $\hat{\theta}$, $\hat{\gamma}$, $\hat{\tau}$, and $\hat{\tau}^{\text{total}}$ alongside, with the AR confidence interval for $\hat{\theta}$.

Scope: As discussed in Appendix F, these robustness values measure sensitivity of the conditional reduced form to confounders W that affect both $\tilde{Z}$ and $\tilde{Y}$. They do *not* assess MCA violations, which operate through ε_T and are absorbed by conditioning on T. Use Step 2 for MCA sensitivity.

4. **Benchmark against observed covariates.** Select one or more covariates X_j that are plausible proxies for the type of confounding under investigation. Compute the relative strength multipliers k_Z and k_Y from (121):

$$k_Z = \frac{R^2_{\tilde{Z}\sim W}}{R^2_{Z\sim X_j|X_{-j},T}}, \qquad k_Y = \frac{R^2_{\tilde{Y}\sim W|\tilde{Z}}}{R^2_{Y\sim X_j|X_{-j},T,Z}}.$$

Under the assumption that the confounder W is orthogonal to X_j after conditioning on the remaining covariates and the instrument (i.e., $R^2_{W\sim X_j|Z,X_{-j}} = 0$), Theorem 6 of Cinelli and Hazlett (2024) translates (k_Z, k_Y) into bounds on the sensitivity parameters $(R^2_{\tilde{Z}\sim W}, R^2_{\tilde{Y}\sim W|Z,X})$ as functions of the observed partial R^2 of X_j. For the outcome, $R^2_{\tilde{Y}\sim W|Z,X} = k_Y\, f^2_{Y\sim X_j|Z,X_{-j}}$, where $f^2_{A\sim B|C} \equiv R^2_{A\sim B|C}/(1-R^2_{A\sim B|C})$ is the partial Cohen's f^2. For the instrument, $R^2_{\tilde{Z}\sim W} \leq \eta(k_Z)\, f^2_{Z\sim X_j|X_{-j}}$, where $\eta(k_Z)$ is a nonlinear function of k_Z and the partial correlation $R_{Z\sim X_j|X_{-j}}$ (see Theorem 6 for the explicit expression). Report bias-adjusted confidence intervals (129) under $k_Z = k_Y = 1$ (confounder as strong as X_j) and under larger multiples (e.g., $k_Z = k_Y = 2$). The table note should include the benchmark-adjusted critical value $t^{\dagger\,\max}_{\alpha,\text{df}-1,\boldsymbol{R}^2}$ under the reference covariate.

5. **Construct sensitivity contour plots.** For each of the lower and upper limits of $\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\theta)$, construct a contour plot with $R^2_{\tilde{Z}\sim W}$ on the horizontal axis and $R^2_{\tilde{Y}\sim W|\tilde{Z}}$ on the vertical axis. At each grid point $(R^2_{\tilde{Z}\sim W}, R^2_{\tilde{Y}\sim W|\tilde{Z}})$, compute the bias-adjusted critical value $t^{\dagger\,\max}_{\alpha,\text{df}-1,\boldsymbol{R}^2}$ from (128) and solve for the boundary of $\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\theta)$ via (129). Mark benchmark covariates as reference points (e.g., red diamonds) on the contour surface. The contour line at zero is particularly informative: it traces the combinations of partial R^2 values at which the confidence interval first includes zero.

Additionally, construct κ-based sensitivity plots as described in the next subsection.

Benchmark Bounds and Sensitivity Plots for the κ-Based Analysis

The κ-based sensitivity analysis requires the researcher to assess the plausibility of hypothesized MCA violations $\kappa_0 = \text{Cov}(\varepsilon_T, \varepsilon_Y)$. This subsection develops formal benchmark bounds—analogous to the R^2-based benchmarks of Cinelli and Hazlett (2024) (Theorem 6)—that calibrate the sensitivity parameter κ against observed covariates X_j.

Proposition 9 (Benchmark Bounds for κ)**.** *Maintain the conditions of Proposition 3 and let X_j be an observed covariate included in the outcome equation (23), with the remaining observed controls denoted X_{-j}. Let $X_j^{\perp}$ denote the residual from projecting X_j on X_{-j}, and let $\varepsilon_T^{(-j)}$ denote the treatment structural error after partialling out Z and X_{-j} but* not *X_j. Define the* leave-one-out treatment-equation

benchmark covariance *for X_j as*

$$\kappa_j \equiv \mathrm{Cov}\big(\varepsilon_T^{(-j)},\, X_j^{\perp}\big). \tag{134}$$

The leave-one-out construction is essential: if X_j were partialled out alongside X_{-j} in defining the residual, OLS orthogonality would force $\widehat{\mathrm{Cov}}(\hat{\varepsilon}_T, X_j) = 0$ mechanically and the benchmark would be vacuous. Then:

(a) *κ_j is consistently estimable by $\hat{\kappa}_j = \widehat{\mathrm{Cov}}(\hat{\varepsilon}_T^{(-j)}, X_j^{\perp})$, where $\hat{\varepsilon}_T^{(-j)}$ and $X_j^{\perp}$ are the OLS leave-one-out residuals defined above.*

(b) *Under the assumption that the unobserved confounder's treatment-equation association is bounded by a multiple $k \geq 0$ of X_j's leave-one-out association, scaled by the ratio of outcome-error to leave-one-out covariate standard deviations (so that k is a unit-free strength parameter comparing equal-variance-normalized associations):*

$$|\kappa| \leq k \cdot |\kappa_j| \cdot \frac{\sigma_{\varepsilon_Y}}{\sigma_{X_j^{\perp}}}, \tag{135}$$

the bias in $\hat{\theta}$ is bounded by

$$|\mathrm{plim}(\hat{\theta}) - \theta| \leq k \cdot \frac{|\kappa_j| \cdot \sigma_{\varepsilon_Y}}{|\rho_{TM}| \cdot \sigma_{X_j^{\perp}}}. \tag{136}$$

(c) *The* benchmark ratio *for covariate X_j,*

$$\widehat{\mathrm{BR}}_j \equiv \frac{|\hat{\kappa}^*|}{|\hat{\kappa}_j| \cdot \hat{\sigma}_{\varepsilon_Y} / \hat{\sigma}_{X_j^{\perp}}}, \tag{137}$$

is consistently estimable and measures the minimum-magnitude *multiple of X_j's leave-one-out treatment-equation association (in outcome-error standard deviation units) needed to overturn the mediator effect. If $\widehat{\mathrm{BR}}_j > 1$, an unobserved confounder with the same treatment-equation association as X_j (and proportional outcome association) is insufficient to reduce $\hat{\theta}$ to zero in magnitude. The direction of confounding required to overturn is given separately by $\mathrm{sign}(\hat{\kappa}^*)$ (see Remark 6).*

(d) *Equivalently, in correlation units, define*

$$\hat{\rho}_{T,X_j} \equiv \frac{\hat{\kappa}_j}{\hat{\sigma}_{\varepsilon_T} \cdot \hat{\sigma}_{X_j^{\perp}}} = \widehat{\mathrm{Corr}}\big(\hat{\varepsilon}_T^{(-j)},\, X_j^{\perp}\big). \tag{138}$$

The breakeven correlation from Remark 6 can be compared directly to $\hat{\rho}_{T,X_j}$: the ratio $|\hat{\rho}^*_{TY}|/|\hat{\rho}_{T,X_j}|$ gives the minimum-magnitude multiple of X_j's treatment-error correlation needed to overturn the result, and equals $\widehat{\mathrm{BR}}_j$.

Proof. Part (a): The leave-one-out residuals $\hat{\varepsilon}_T^{(-j)}$ and $X_j^{\perp}$ are obtained by OLS projections onto $\{Z, X_{-j}\}$ and X_{-j} respectively. By OLS consistency under the maintained moment conditions, $\hat{\varepsilon}_T^{(-j)} \xrightarrow{p} \varepsilon_T^{(-j)}$

and $X_j^\perp \xrightarrow{p}$ its population counterpart. By the continuous mapping theorem, $\hat{\kappa}_j = \widehat{\text{Cov}}(\hat{\varepsilon}_T^{(-j)}, X_j^\perp) \xrightarrow{p} \text{Cov}(\varepsilon_T^{(-j)}, X_j^\perp) = \kappa_j$.

Part (b): The bias formula (42) gives $|\text{plim}(\hat{\theta}) - \theta| = |\kappa|/|\rho_{TM}|$. Substituting the calibration bound (135) yields (136).

Part (c): The signed breakeven $\hat{\kappa}^*$ converges to $\kappa^* = \theta\rho_{TM}$ under MCA (Proposition 3(c)). By part (a), $\hat{\kappa}_j \xrightarrow{p} \kappa_j$, and $\hat{\sigma}_{\varepsilon_Y}/\hat{\sigma}_{X_j^\perp} \xrightarrow{p} \sigma_{\varepsilon_Y}/\sigma_{X_j^\perp}$ by the CMT. Taking absolute values term-by-term, $\widehat{\text{BR}}_j \xrightarrow{p} |\kappa^*|/(|\kappa_j| \cdot \sigma_{\varepsilon_Y}/\sigma_{X_j^\perp})$. The interpretation follows: $\hat{\theta}$ is overturned in magnitude when $|\kappa| \geq |\hat{\kappa}^*|$, i.e., when $k \geq \widehat{\text{BR}}_j$ in (135).

Part (d): From the definitions, $|\hat{\rho}_{TY}^*|/|\hat{\rho}_{T,X_j}| = [|\hat{\kappa}^*|/(\hat{\sigma}_{\varepsilon_T}\hat{\sigma}_{\varepsilon_Y})]/[|\hat{\kappa}_j|/(\hat{\sigma}_{\varepsilon_T}\hat{\sigma}_{X_j^\perp})] = |\hat{\kappa}^*| \cdot \hat{\sigma}_{X_j^\perp}/(|\hat{\kappa}_j| \cdot \hat{\sigma}_{\varepsilon_Y}) = \widehat{\text{BR}}_j$ by (137). □

Comparison with R^2-Based Benchmarks. The κ-based benchmark bounds differ from the R^2-based benchmarks of Cinelli and Hazlett (2024) in two respects. First, the κ parameterization is *one-dimensional*: the researcher specifies only the treatment-equation association of the confounder (via κ_j), rather than the two-dimensional (k_Z, k_Y) pair required by the R^2 framework. This simplification arises because MCA violations operate through a single channel ($\text{Cov}(\varepsilon_T, \varepsilon_Y)$), whereas reduced-form confounders affect both the instrument and the outcome. Second, the κ-based benchmarks target MCA specifically, while the R^2-based benchmarks target reduced-form confounders (Appendix F). Both types of benchmarks should be reported when available.

Reporting Format. The benchmark analysis is reported as a table with columns: benchmark covariate X_j; leave-one-out treatment-equation covariance $\hat{\kappa}_j = \widehat{\text{Cov}}(\hat{\varepsilon}_T^{(-j)}, X_j^\perp)$; leave-one-out treatment-error correlation $\hat{\rho}_{T,X_j} = \widehat{\text{Corr}}(\hat{\varepsilon}_T^{(-j)}, X_j^\perp)$; benchmark ratio $\widehat{\text{BR}}_j$; and the κ-adjusted estimate $\hat{\theta}_{\text{adj}}(\hat{\kappa}_j \cdot \hat{\sigma}_{\varepsilon_Y}/\hat{\sigma}_{X_j^\perp})$ under the assumption that the confounder is as strong as X_j (i.e., $k = 1$). A benchmark ratio $\widehat{\text{BR}}_j \gg 1$ for all observed covariates indicates that the mediator effect is robust to MCA violations of the type captured by the observed covariates.

Caveat: Pre-Determined Covariates. When all observed covariates X_j are pre-determined or exogenous by design, the leave-one-out covariance $\text{Cov}(\varepsilon_T^{(-j)}, X_j^\perp) \approx 0$ as a substantive matter (independent of any partialling artifact), and the benchmark ratios $\widehat{\text{BR}}_j$ are undefined or arbitrarily large. In this regime—which is typical in well-designed empirical studies—the benchmark framework is uninformative, and the researcher should assess the plausibility of MCA violations directly using the breakeven correlation $\hat{\rho}_{TY}^*$ (Remark 6), the LR estimate $\hat{\kappa}$ when $K \geq 2$ (Section 5.1), or domain-specific reasoning about the magnitude of treatment–outcome confounding conditional on the mediator channel. The benchmark bounds are most useful when the covariate set includes variables whose exogeneity is not assumed *a priori*—for instance, post-treatment covariates or variables measured with error—so that $\kappa_j \neq 0$ provides a non-trivial calibration target.

κ-Based Sensitivity Plots. Two complementary visualizations display the sensitivity of $\hat{\theta}$ to MCA violations.

Plot 1 (One-dimensional): Plot the κ-adjusted estimate $\hat{\theta}_{\text{adj}}(\kappa_0) = \hat{\theta} - \kappa_0/\widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$ as a function of κ_0 on the horizontal axis. The function is linear with slope $-1/\widehat{\text{Cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$. Include: (i) a shaded

$1-\alpha$ confidence band $\hat{\theta}_{\text{adj}}(\kappa_0) \pm z_{\alpha/2} \cdot \widehat{\text{SE}}(\hat{\theta})$; (ii) a horizontal dashed line at zero; (iii) vertical lines at $\kappa_0 = 0$ (MCA), $\kappa_0 = \hat{\kappa}^*$ (point-estimate breakeven), and $\kappa_0 = \hat{\kappa}^*_{\text{CI}}$ (inference breakeven); (iv) a secondary horizontal axis showing $\rho_{TY} = \kappa_0/(\hat{\sigma}_{\varepsilon_T}\hat{\sigma}_{\varepsilon_Y})$; and (v) benchmark diamonds at $\kappa_0 = \hat{\kappa}_j \cdot \hat{\sigma}_{\varepsilon_Y}/\hat{\sigma}_{X_j^\perp}$ for each observed covariate X_j (using the leave-one-out covariance and standard deviation defined in Proposition 9). When $K \geq 2$, add a vertical line at $\kappa_0 = \hat{\kappa}$ (LR estimate) with its confidence interval. This plot directly communicates: "the mediator effect remains significant as long as κ stays below $\hat{\kappa}^*_{\text{CI}}$."

Plot 2 (Two-dimensional, compound sensitivity): Plot contour lines of $\hat{\theta}_{\text{adj}}$ in the space $(\rho_{TM}, \rho_{TY}) \in [-1,1]^2$, where $\rho_{TM} = \text{Corr}(\varepsilon_T, \varepsilon_M)$ and $\rho_{TY} = \text{Corr}(\varepsilon_T, \varepsilon_Y)$. In the unit-variance simplification ($\sigma_{\varepsilon_T} = \sigma_{\varepsilon_M} = \sigma_{\varepsilon_Y} = 1$), the bias equals ρ_{TY}/ρ_{TM}, so contour lines of the adjusted estimate $\theta_{\text{adj}} = \hat{\theta} - \rho_{TY}/\rho_{TM}$ are hyperbolas. Include: (i) shading for the region where $\hat{\theta}_{\text{adj}} > 0$ and significant (green), the region where the CI includes zero (yellow), and the region where $\hat{\theta}_{\text{adj}} < 0$ (red); (ii) a diamond at $(\hat{\rho}_{TM}, 0)$ marking the MCA-consistent estimate; (iii) the zero contour $\rho_{TY}/\rho_{TM} = \hat{\theta}$ as a dashed curve; and (iv) the positive-definiteness boundary of the $(\varepsilon_T, \varepsilon_M, \varepsilon_Y)$ correlation matrix (the feasibility region is $\det(\Sigma) > 0$ given $\rho_{MY} = \widehat{\text{Corr}}(\hat{\varepsilon}_M, \hat{\varepsilon}_Y)$). This plot addresses compound sensitivity: what happens if both the collider strength (ρ_{TM}) differs from its estimate and there is an MCA violation ($\rho_{TY} \neq 0$)?

Specification Tests in Designs with an Independent Mediator Instrument

This subsection provides the formulas and asymptotic distributions for the J-test and the targeted t-test referenced in Section 5.1. These tests are diagnostic for MCA only in the Frölich–Huber two-instrument design, where the instrument vector $\boldsymbol{Z}_i$ contains a component with a direct effect on the mediator; in the canonical single-instrument design they have zero asymptotic power against MCA violations (see the observational-equivalence argument of Section 5.1).

J-Test (Overall Specification). The Hansen J-statistic tests whether all moment conditions in (31) hold jointly.[29] Let $\hat{\boldsymbol{\delta}}$ denote the efficient two-step GMM estimator with first-step weighting matrix $\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{g},n}$. Under correct specification:

$$\hat{J} = n\, \bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\delta}})' \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{g},n} \bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\delta}}) \xrightarrow{d} \chi^2_{2(K-1)}. \tag{139}$$

Large values indicate that one or more moment conditions fail, but the test does not identify which condition is violated.

t-Test (Mediated Confounding Moment). To assess the specific moment condition corresponding to MCA, we compute the t-ratio of the evaluated fourth moment. Under correct specification, the evaluated moments $\sqrt{n}\bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\delta}})$ are asymptotically normal with covariance $\boldsymbol{\Sigma}_{\boldsymbol{g}} - \boldsymbol{G}(\boldsymbol{G}'\boldsymbol{\Sigma}_{\boldsymbol{g}}^{-1}\boldsymbol{G})^{-1}\boldsymbol{G}'$. The t-ratio for the k-th moment is:

$$t_k = \frac{\left[\bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\delta}})\right]_k}{\left[\text{SE}\left(\bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\delta}})\right)\right]_k} \xrightarrow{d} \text{N}(0,1), \quad \text{where} \quad \text{SE}\left(\bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\delta}})\right) = \text{diag}\left(\frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{g},n} - \boldsymbol{G}_n\left(\boldsymbol{G}_n'\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{g},n}\boldsymbol{G}_n\right)^{-1}\boldsymbol{G}_n'}{n}\right)^{1/2} \tag{140}$$

[29] The J-statistic is zero for just-identified models ($K = 1$).

and $\boldsymbol{G}_n = \partial \bar{\boldsymbol{g}}_n(\boldsymbol{\delta})/\partial \boldsymbol{\delta}'\big|_{\hat{\boldsymbol{\delta}}}$ is the sample Jacobian. Applying this test to the last moment condition targets the MCA specifically.

G Proofs and Supplementary Material for Section 6

G.1 Proofs

Proof of Proposition 4. Identification of $\mathrm{E}(M(1) \mid U_T = p)$. By mediation exclusion, $M(t,z) = M(t)$, and by consistency, $M = M(T)$. Combining with $T = \mathbf{1}[P_T(Z) \geq U_T]$ from (51),

$$M \cdot T \;=\; M(T) \cdot T \;=\; M(1) \cdot T \;=\; M(1) \cdot \mathbf{1}[P_T(Z) \geq U_T],$$

where the second equality uses $T \in \{0,1\}$. Conditioning on $P_T(Z) = p$ and using $(M(1), U_T) \perp\!\!\!\perp P_T(Z)$—which holds because $P_T(Z)$ is a function of Z, while $(M(1), U_T)$ is a function of (ν_T, ν_Y) alone, and structural IV exogeneity gives $Z \perp\!\!\!\perp (\nu_T, \nu_Y)$—

$$\mathrm{E}(M \cdot T \mid P_T(Z) = p) \;=\; \mathrm{E}\left(M(1) \cdot \mathbf{1}[p \geq U_T]\right) \;=\; \int_0^p \mathrm{E}\left(M(1) \mid U_T = u\right) du,$$

where the second equality uses $U_T \sim \mathrm{Uniform}[0,1]$. By the fundamental theorem of calculus, the derivative in p equals $\mathrm{E}(M(1) \mid U_T = p)$ at every point of continuity, establishing (52).

Identification of $\mathrm{E}(M(0) \mid U_T = p)$. By the same chain, $M \cdot (1 - T) = M(0) \cdot \mathbf{1}[U_T > P_T(Z)]$. Conditioning and integrating,

$$\mathrm{E}(M \cdot (1-T) \mid P_T(Z) = p) \;=\; \int_p^1 \mathrm{E}\left(M(0) \mid U_T = u\right) du.$$

Differentiating in p (the lower limit of integration) gives $-\mathrm{E}(M(0) \mid U_T = p)$, establishing (53). □

Proof of Corollary 2. Identification of $\rho(1,p)$. By Proposition 4, $\partial\,\mathrm{E}(M \cdot T \mid P_T(Z) = p)/\partial p = \mathrm{E}(M(1) \mid U_T = p)$. Under mediator separability (55), $M(1) = \mathbf{1}[\rho(1, U_T) \geq U_Y]$, so

$$\mathrm{E}(M(1) \mid U_T = p) \;=\; \mathrm{E}\left(\mathbf{1}[\rho(1,p) \geq U_Y] \,\middle|\, U_T = p\right) \;=\; \Pr\left(U_Y \leq \rho(1,p)\right) \;=\; \rho(1,p),$$

where the second equality uses $U_T \perp\!\!\!\perp U_Y$ (MCA) and the third uses $U_Y \sim \mathrm{Uniform}[0,1]$.

Identification of $\rho(0,p)$. By Proposition 4, $\partial\,\mathrm{E}(M \cdot (1-T) \mid P_T(Z) = p)/\partial p = -\mathrm{E}(M(0) \mid U_T = p)$. By the same chain with $M(0) = \mathbf{1}[\rho(0, U_T) \geq U_Y]$, MCA, and uniformity of U_Y, $\mathrm{E}(M(0) \mid U_T = p) = \rho(0,p)$, giving the second formula. □

Proof of Theorem 2. By consistency, $Y = Y(T, M)$ and $M = M(T)$. We establish (62) in detail for $(t,m) = (1,1)$ and $(1,0)$; the cases with $t = 0$ follow by parallel arguments using (53), with the lower limit of integration shifting to p and contributing the additional $(-1)^{1-t}$ sign factor.

Step 1 (first derivative for $(t,m) = (1,1)$*).* Using consistency, $T \in \{0,1\}$, and $M(1) \in \{0,1\}$,

$$Y \cdot M \cdot T \;=\; Y(T, M(T)) \cdot M(T) \cdot T \;=\; Y(1, M(1)) \cdot M(1) \cdot T \;=\; Y(1,1) \cdot M(1) \cdot \mathbf{1}[P_T(Z) \geq U_T],$$

where the third equality uses $Y(1, M(1)) \cdot M(1) = Y(1,1) \cdot M(1)$ for binary $M(1)$ (the $M(1) = 0$ case contributes zero) and $T = \mathbf{1}[P_T(Z) \geq U_T]$ from (51). Conditioning on $P_T(Z) = p$ and using structural IV exogeneity $Z \perp\!\!\!\perp (\nu_T, \nu_Y, \epsilon)$ (so $P_T(Z)$, a function of Z, is independent of every function

of (ν_T, ν_Y, ϵ), including $(Y(1,1), M(1), U_T)$),

$$\mathrm{E}(Y \cdot M \cdot T \mid P_T(Z) = p) \;=\; \mathrm{E}\left(Y(1,1) \cdot M(1) \cdot \mathbf{1}[p \geq U_T]\right) \;=\; \int_0^p \mathrm{E}\left(Y(1,1) \cdot M(1) \mid U_T = u\right) du,$$

using $U_T \sim \mathrm{Uniform}[0,1]$. By the fundamental theorem of calculus,

$$\frac{\partial\, \mathrm{E}(Y \cdot M \cdot T \mid P_T(Z) = p)}{\partial p} = \mathrm{E}(Y(1,1) \cdot M(1) \mid U_T = p). \tag{141}$$

Step 2 (substituting the mediation equation). Using $M(1) = \mathbf{1}[\rho(1, U_T) \geq U_Y]$ from (55):

$$\begin{aligned} \mathrm{E}(Y(1,1) \cdot M(1) \mid U_T = p) &= \mathrm{E}\big(Y(1,1) \cdot \mathbf{1}[\rho(1,p) \geq U_Y] \,\big|\, U_T = p\big) \\ &= \mathrm{E}\big(Y(1,1) \cdot \mathbf{1}[\rho(1,p) \geq U_Y]\big), \end{aligned} \tag{142}$$

where the last equality uses $U_T \perp\!\!\!\perp (Y(t,m), U_Y)$ from (49).

Step 3 (second derivative). Since $U_Y \sim \mathrm{Uniform}[0,1]$, the right-hand side of (142) can be written as

$$\mathrm{E}\big(Y(1,1) \cdot \mathbf{1}[\rho(1,p) \geq U_Y]\big) = \int_0^{\rho(1,p)} \mathrm{E}(Y(1,1) \mid U_Y = u)\, du. \tag{143}$$

By Assumption 2(iii)–(iv), $\rho(1,\cdot)$ is continuously differentiable with $\rho'(1,p) \neq 0$, and the latent conditional mean $\mu_{11}(u) = \mathrm{E}(Y(1,1) \mid U_Y = u)$ has a bounded continuous version on the image $\rho(1, (p_L, p_H))$. By the fundamental theorem of calculus (applied to the continuous integrand μ_{11}) and the chain rule,

$$\frac{\partial}{\partial p} \int_0^{\rho(1,p)} \mathrm{E}(Y(1,1) \mid U_Y = u)\, du \;=\; \mathrm{E}(Y(1,1) \mid U_Y = \rho(1,p)) \cdot \rho'(1,p),$$

where the right-hand side is well-defined pointwise because μ_{11} is continuous at $\rho(1,p)$. Combining with (143) yields

$$\frac{\partial\, \mathrm{E}(Y(1,1) \cdot M(1) \mid U_T = p)}{\partial p} = \rho'(1,p) \cdot \mathrm{E}(Y(1,1) \mid U_Y = \rho(1,p)), \tag{144}$$

with $\rho'(1,p) \neq 0$ by Assumption 2(iii). Combining (141) and (144) establishes (62) for $(t,m) = (1,1)$, where the sign factor $(-1)^{1-1} \cdot (-1)^{1-1} = 1$.

Case $(t,m) = (1,0)$. By consistency and the binary structure,

$$Y \cdot \mathbf{1}[M = 0] \cdot T \;=\; Y(1,0) \cdot (1 - M(1)) \cdot T \;=\; Y(1,0) \cdot \mathbf{1}[U_Y > \rho(1, U_T)] \cdot \mathbf{1}[p \geq U_T].$$

Conditioning, applying (49), and using $U_Y \sim \mathrm{Uniform}[0,1]$,

$$\mathrm{E}(Y \cdot \mathbf{1}[M = 0] \cdot T \mid P_T(Z) = p) \;=\; \int_0^p \int_{\rho(1,u)}^1 \mathrm{E}(Y(1,0) \mid U_Y = v)\, dv\, du.$$

Differentiating twice in p (Leibniz rule on the outer upper limit, then chain rule on the lower limit of the

resulting integral):

$$\frac{\partial^2 \mathrm{E}(Y \cdot \mathbf{1}[M=0] \cdot T \mid P_T(Z)=p)}{\partial p^2} \;=\; -\rho'(1,p) \cdot \mathrm{E}(Y(1,0) \mid U_Y = \rho(1,p)),$$

matching (62) with sign factor $(-1)^{1-0} \cdot (-1)^{1-1} = -1$. The factor $(-1)^{1-m}$ thus reflects whether the indicator $\mathbf{1}[M=m]$ selects $\{U_Y \le \rho\}$ or $\{U_Y > \rho\}$.

Cases with $t=0$. By the parallel argument applied to $\mathbf{1}[T=0]$, using (53) in place of (52), the integration over U_T runs from p to 1 rather than from 0 to p; differentiating in the lower limit contributes the additional (-1) that produces the $(-1)^{1-t}$ factor. The combined sign $(-1)^{1-m} \cdot (-1)^{1-t}$ in (62) is verified for all four cases.

Step 4 (the MME and the ratio form). Summing (62) over $m \in \{0,1\}$ yields

$$\frac{\partial^2 \mathrm{E}(Y \cdot T \mid P_T(Z)=p)}{\partial p^2} = \rho'(1,p)\,\{-\mathrm{E}(Y(1,0) \mid U_Y = u) + \mathrm{E}(Y(1,1) \mid U_Y = u)\} = \rho'(1,p)\Delta_M(1,u),$$

where $u = \rho(1,p)$. This proves (65) for $t=1$; the sign $(-1)^{1-t}$ gives the corresponding formula for $t=0$.

Next consider the treatment-stratum mediator expectation

$$H_t(p) = \mathrm{E}(M \cdot \mathbf{1}[T=t] \mid P_T(Z)=p).$$

By Proposition 4 and Corollary 2,

$$H_t'(p) = (-1)^{1-t}\rho(t,p), \qquad H_t''(p) = (-1)^{1-t}\rho'(t,p).$$

Substituting into (65) gives

$$\Delta_M(t,u) = \frac{\partial^2 \mathrm{E}(Y \cdot \mathbf{1}[T=t] \mid P_T(Z)=p)/\partial p^2}{\partial^2 \mathrm{E}(M \cdot \mathbf{1}[T=t] \mid P_T(Z)=p)/\partial p^2} = \frac{G_t''(p)}{H_t''(p)},$$

which is (64). □

Proof of Proposition 5. Key factorization. Conditional on $U_Y = u$, $Y(t,1) - Y(t,0)$ is a function of (ν_Y, ϵ) and $\mathbf{1}[\rho(t', U_T) \ge u]$ is a function of U_T. By MCA, $U_T \perp\!\!\!\perp (\nu_Y, \epsilon)$, so the two are conditionally independent given U_Y. Combined with $U_T \perp\!\!\!\perp U_Y$ and $U_T \sim \mathrm{Uniform}[0,1]$,

$$\mathrm{E}\left[(Y(t,1) - Y(t,0)) \cdot \mathbf{1}[\rho(t', U_T) \ge u] \,\middle|\, U_Y = u\right] \;=\; \Delta_M(t,u) \cdot S_M(t',u). \tag{145}$$

Part (i). By the tower property and $U_Y \sim \mathrm{Uniform}[0,1]$,

$$\mathrm{E}[Y(t,1) - Y(t,0)] \;=\; \int_0^1 \mathrm{E}[Y(t,1) - Y(t,0) \mid U_Y = u]\, du \;=\; \int_0^1 \Delta_M(t,u)\, du.$$

Part (ii). Apply (145) with $t' = t$, integrate over U_Y, and divide by $\Pr(M(t)=1) = \int_0^1 S_M(t,v)\, dv$ (which follows from $\mathrm{E}(M(t)) = \mathrm{E}[\mathrm{E}(M(t) \mid U_Y)]$ and $U_Y \sim \mathrm{Uniform}$).

Part (iii). Using $\mathbf{1}[M(t)=0] = 1 - \mathbf{1}[M(t)=1]$ and parts (i)–(ii),

$$\mathrm{E}[(Y(t,1)-Y(t,0))\cdot\mathbf{1}[M(t)=0]] \;=\; \int_0^1 \Delta_M(t,u)\,du - \int_0^1 \Delta_M(t,u)\,S_M(t,u)\,du \;=\; \int_0^1 \Delta_M(t,u)\big(1-S_M(t,u)\big)\,du$$

Dividing by $\Pr(M(t)=0) = \int_0^1 (1-S_M(t,v))\,dv$ gives the integral form in (68).

Part (iv). Since $M(t') \in \{0,1\}$,

$$Y(t,M(t')) \;=\; Y(t,0)\cdot\mathbf{1}[M(t')=0] + Y(t,1)\cdot\mathbf{1}[M(t')=1] \;=\; Y(t,0) + (Y(t,1)-Y(t,0))\cdot M(t'),$$

so $\mathrm{NIE}(t) = \mathrm{E}[(Y(t,1)-Y(t,0))\cdot(M(1)-M(0))]$. Applying (145) with $t'=1$ and $t'=0$, taking the difference, and integrating over U_Y,

$$\mathrm{NIE}(t) \;=\; \int_0^1 \Delta_M(t,u)\big[S_M(1,u) - S_M(0,u)\big]\,du.$$

This argument uses only the binary-mediator structure and MCA; it does not invoke outcome separability.

Part (v). Since $M(t) \in \{0,1\}$, for each $a \in \{0,1\}$,

$$Y(a,M(t)) \;=\; Y(a,0) + \big(Y(a,1) - Y(a,0)\big)\cdot M(t).$$

Applying this identity with $a=1$ and $a=0$, taking the difference, and then taking expectations,

$$\begin{aligned}\mathrm{NDE}(t) = \mathrm{E}[Y(1,0) - Y(0,0)] \\ + \mathrm{E}\big[\big(Y(1,1)-Y(1,0)\big)\cdot M(t)\big] - \mathrm{E}\big[\big(Y(0,1)-Y(0,0)\big)\cdot M(t)\big]. \end{aligned} \tag{146}$$

For each $a \in \{0,1\}$, apply (145) with outcome treatment index a and counterfactual mediator index t:

$$\mathrm{E}\big[\big(Y(a,1)-Y(a,0)\big)\cdot M(t) \mid U_Y = u\big] \;=\; \Delta_M(a,u)\cdot S_M(t,u).$$

Integrating over $U_Y \sim \mathrm{Uniform}[0,1]$ gives

$$\mathrm{E}\big[\big(Y(a,1)-Y(a,0)\big)\cdot M(t)\big] \;=\; \int_0^1 \Delta_M(a,u) S_M(t,u)\,du.$$

Substituting this expression into (146) for $a=1$ and $a=0$ yields

$$\mathrm{NDE}(t) \;=\; \mathrm{E}[Y(1,0)-Y(0,0)] + \int_0^1 \big[\Delta_M(1,u) - \Delta_M(0,u)\big] S_M(t,u)\,du.$$

This argument uses the binary-mediator expansion and the same MCA-based factorization as the NIE representation; it does not invoke outcome separability. □

Proof of Lemma 1. Under IV exogeneity and MCA,

$$(Z,\nu_T) \perp\!\!\!\perp (\nu_Y,\epsilon).$$

Since $T = f_T(Z, \nu_T)$ and $S_T = (T(z_0), T(z_1))$ are functions of (Z, ν_T), whereas $Y(t, m) = f_Y(t, m, \nu_Y, \epsilon)$ is a function of (ν_Y, ϵ), we have

$$Y(t, m) \perp\!\!\!\perp (S_T, T).$$

Weak union therefore yields

$$Y(t, m) \perp\!\!\!\perp S_T \mid T.$$

□

Proof of Theorem 3. We prove the result for $t = 0$; the case $t = 1$ follows by parallel argument (sketched at the end).

Step 1 (Mixture representation of stratum means). Under treatment monotonicity (Assumption 3), the $T = 0$ stratum contains N-types at any Z and C-types only when $Z = z_0$ (since C-types have $T(z_1) = 1$). Hence $\Pr(T = 0 \mid Z = z_0) = \pi_N + \pi_C = 1 - p_0$ and $\Pr(T = 0 \mid Z = z_1) = \pi_N = 1 - p_1$, giving the type shares

$$\Pr(N \mid T = 0, z_0) = \tfrac{\pi_N}{\pi_N + \pi_C}, \qquad \Pr(C \mid T = 0, z_0) = \tfrac{\pi_C}{\pi_N + \pi_C}, \qquad \Pr(N \mid T = 0, z_1) = 1.$$

By consistency, on the event $\{T = 0\}$, $Y = Y(0, M(0))$. Therefore

$$\mathrm{E}[Y \mid z_0, T{=}0] = \tfrac{\pi_N}{\pi_N + \pi_C}\, \mathrm{E}[Y(0, M(0)) \mid N] + \tfrac{\pi_C}{\pi_N + \pi_C}\, \mathrm{E}[Y(0, M(0)) \mid C], \tag{147}$$

$$\mathrm{E}[Y \mid z_1, T{=}0] = \mathrm{E}[Y(0, M(0)) \mid N]. \tag{148}$$

Subtracting:

$$\mathrm{E}[Y \mid z_0, T{=}0] - \mathrm{E}[Y \mid z_1, T{=}0] = \tfrac{\pi_C}{\pi_N + \pi_C}\big(\mathrm{E}[Y(0, M(0)) \mid C] - \mathrm{E}[Y(0, M(0)) \mid N]\big). \tag{149}$$

The analogous identity holds with M in place of Y. The type-share factor $\pi_C/(\pi_N + \pi_C)$ cancels in the ratio, giving

$$\theta^W(0) = \frac{\mathrm{E}[Y(0, M(0)) \mid C] - \mathrm{E}[Y(0, M(0)) \mid N]}{\mathrm{E}[M(0) \mid C] - \mathrm{E}[M(0) \mid N]}. \tag{150}$$

Step 2 (Decomposition via conditional exogeneity). Since $M(0) \in \{0, 1\}$,

$$Y(0, M(0)) = Y(0, 0) + (Y(0, 1) - Y(0, 0)) \cdot M(0).$$

For each type $s \in \{N, C\}$,

$$\mathrm{E}[Y(0, M(0)) \mid s] = \mathrm{E}[Y(0, 0) \mid s] + \mathrm{E}[(Y(0, 1) - Y(0, 0)) \cdot M(0) \mid s].$$

The type contrast in (150) is generated inside the $T = 0$ stratum. Lemma 1 implies that, within this stratum, the adjacent treatment-type composition is independent of $Y(0, 0)$. Since the N component and the C component appearing in the contrast are both evaluated with treatment held at 0, the baseline terms coincide and cancel in (150):

$$\theta^W(0) = \frac{\mathrm{E}[(Y(0, 1) - Y(0, 0)) \cdot M(0) \mid C] - \mathrm{E}[(Y(0, 1) - Y(0, 0)) \cdot M(0) \mid N]}{\mathrm{E}[M(0) \mid C] - \mathrm{E}[M(0) \mid N]}. \tag{151}$$

Step 3 (Factoring the cross term). For each type s, condition on U_T in the type-s block and on $U_Y = v$. Under MCA, $\nu_T \perp\!\!\!\perp (\nu_Y, \epsilon)$, hence $U_T \perp\!\!\!\perp U_Y$ and (within the level set $\{U_Y = v\}$) $\mathrm{E}[Y(0,1) - Y(0,0) \mid U_Y = v] = \Delta_M(0,v)$ as in Definition 1. Using $M(0) = \mathbf{1}[\rho(0,U_T) \geq U_Y]$ and the conditional independence above,

$$\begin{aligned}\mathrm{E}[(Y(0,1) - Y(0,0)) \cdot M(0) \mid s] &= \mathrm{E}_{U_T|s}\left[\int_0^1 \Delta_M(0,v) \cdot \mathbf{1}[\rho(0,U_T) \geq v]\, dv\right] \\ &= \mathrm{E}_{U_T|s}\left[\int_0^{\rho(0,U_T)} \Delta_M(0,v)\, dv\right], \end{aligned} \tag{152}$$

using $U_Y \sim \mathrm{Uniform}[0,1]$ for the inner integral.

Define $F(u_T) \equiv \int_0^{\rho(0,u_T)} \Delta_M(0,v)\, dv$. Then $\mathrm{E}[(Y(0,1) - Y(0,0)) \cdot M(0) \mid s] = \mathrm{E}[F(U_T) \mid s]$, and (151) becomes

$$\theta^W(0) = \frac{\mathrm{E}[F(U_T) \mid C] - \mathrm{E}[F(U_T) \mid N]}{\mathrm{E}[\rho(0,U_T) \mid C] - \mathrm{E}[\rho(0,U_T) \mid N]}, \tag{153}$$

where we used $\mathrm{E}[M(0) \mid s] = \mathrm{E}[\rho(0,U_T) \mid s]$ (from $\mathrm{E}[\mathbf{1}[\rho(0,U_T) \geq U_Y] \mid U_T] = \rho(0,U_T)$ by $U_Y \sim \mathrm{Uniform}$ and MCA).

Step 4 (Survival-function weights via Fubini). Since $U_T \mid (S_T = s)$ has known support on the type-s block,

$$\begin{aligned}\mathrm{E}[F(U_T) \mid s] &= \mathrm{E}_{U_T|s}\left[\int_0^1 \Delta_M(0,v) \cdot \mathbf{1}[v \leq \rho(0,U_T)]\, dv\right] \\ &= \int_0^1 \Delta_M(0,v) \cdot \Pr\big(\rho(0,U_T) \geq v \;\big|\; S_T = s\big)\, dv = \int_0^1 \Delta_M(0,v)\, S_s(v)\, dv, \end{aligned} \tag{154}$$

where $S_s(v) \equiv \Pr(\rho(0,U_T) \geq v \mid S_T = s)$ and the exchange of integral and expectation is by Fubini ($\Delta_M(0,\cdot)$ integrable by Assumption 2). Similarly $\mathrm{E}[\rho(0,U_T) \mid s] = \int_0^1 S_s(v)\, dv$ (mean equals integral of survival function for a nonnegative variable). Substituting into (153):

$$\theta^W(0) = \frac{\int_0^1 \Delta_M(0,v)\,[S_C(v) - S_N(v)]\, dv}{\int_0^1 [S_C(v) - S_N(v)]\, dv}. \tag{155}$$

Under compositional mediator monotonicity (Assumption 4), $S_C(v) \geq S_N(v)$ for all v, so the weight $\omega_0(v) = S_C(v) - S_N(v)$ is nonnegative. This establishes the survival-weighted representation of Proposition 6 for $t = 0$.

Step 5 (Closing $\theta^W(0) = \theta^{LM}(0)$). By definition $\theta^{LM}(0) \equiv \mathrm{E}[Y(0,1) - Y(0,0) \mid \mathcal{C}_0^M]$. Conditional on U_Y, the type-side resistance variation that determines $\mathcal{C}_0^M$ is independent of the outcome-side variation in $Y(0,1) - Y(0,0)$ by MCA. Hence

$$\theta^{LM}(0) = \mathrm{E}\big[\mathrm{E}[Y(0,1) - Y(0,0) \mid U_Y, \mathcal{C}_0^M] \,\big|\, \mathcal{C}_0^M\big] = \mathrm{E}[\Delta_M(0,U_Y) \mid \mathcal{C}_0^M].$$

By Bayes' rule and $U_Y \sim \mathrm{Uniform}[0,1]$, $f_{U_Y|\mathcal{C}_0^M}(v) = \Pr(\mathcal{C}_0^M \mid U_Y = v)/\Pr(\mathcal{C}_0^M)$. Under compositional mediator monotonicity, $\{M(0,N) = 1\} \subseteq \{M(0,C) = 1\}$, so

$$\Pr(\mathcal{C}_0^M \mid U_Y = v) = \Pr(M(0,C) = 1 \mid U_Y = v) - \Pr(M(0,N) = 1 \mid U_Y = v) = S_C(v) - S_N(v),$$

and $\Pr(\mathcal{C}_0^M) = \int_0^1 [S_C(u) - S_N(u)]\,du$. Substituting,

$$\theta^{LM}(0) = \int_0^1 \Delta_M(0,v) \cdot \frac{S_C(v) - S_N(v)}{\int_0^1 [S_C(u) - S_N(u)]\,du}\,dv = \theta^W(0)$$

by Step 4, establishing $\theta^W(0) = \theta^{LM}(0)$.

Case $t = 1$. The $T = 1$ stratum contains A-types at any Z and C-types only when $Z = z_1$. By the parallel argument with $(s_H, s_L) = (A, C)$ and Assumption 4's $M(1, A) \geq M(1, C)$,

$$\theta^W(1) = \frac{\int_0^1 \Delta_M(1,v)[S_A(v) - S_C(v)]\,dv}{\int_0^1 [S_A(v) - S_C(v)]\,dv}, \qquad \Pr(\mathcal{C}_1^M \mid U_Y = v) = S_A(v) - S_C(v),$$

and $\theta^W(1) = \theta^{LM}(1)$. □

Proof of Proposition 6. The survival-weighted representation (84)–(85) is established in Step 4 of the proof of Theorem 3 (equation (155) for $t = 0$, with the parallel argument using (S_A, S_C) for $t = 1$). The identification $\theta^W(t) = \theta^{LM}(t)$—i.e., the equivalence between the survival-weighted form and the conditional expectation $\mathrm{E}[Y(t,1) - Y(t,0) \mid \mathcal{C}_t^M]$—is the content of Step 5 of the same proof.

Nonnegativity of ω_t under compositional mediator monotonicity is established alongside Step 4. The normalizing-constant identity $\int_0^1 \omega_t(v)\,dv = \mathrm{E}[\rho(t, U_T) \mid s_H] - \mathrm{E}[\rho(t, U_T) \mid s_L]$ follows from the standard "mean equals integral of survival function" identity for nonnegative random variables (used in the same Step 4). For convergence to the pointwise MME as $p_1 - p_0 \to 0$, embed the binary instrument in a sequence with $p_1^{(n)} - p_0^{(n)} \to 0$ and $p_0^{(n)} \to p$ for some interior $p \in (p_L, p_H)$. As the type blocks shrink, the survival functions $S_{s_H}^{(n)}$ and $S_{s_L}^{(n)}$ concentrate around $v = \rho(t,p)$, so the weight $\omega_t^{(n)}$ approaches a point mass at $\rho(t,p)$. By continuity of $\Delta_M(t,\cdot)$ (Assumption 2), $\theta^{LM}(t) \to \Delta_M(t, \rho(t,p))$. □

G.2 Type Structure under Alternative Mediator Restrictions

This subsection illustrates the difference between standard mediator monotonicity and mediator separability using a type-based representation. The goal is to clarify which margins of variation are exploited by the identification results in Section 6.

Under treatment monotonicity (Assumption 3), the treatment-type variable satisfies $S_T \in \{N, C, A\}$ almost surely. Standard mediator monotonicity imposes an ordering across treatment states, $M(1) \geq M(0)$, which yields a seven-type partition. By contrast, mediator separability (Section 6.2) allows the MMR curves $\rho(1,\cdot)$ and $\rho(0,\cdot)$ to vary independently across the latent treatment resistance, potentially crossing and generating an additional type (mediator defiers), resulting in eight types.

Figure G.1 visualizes the two environments. Panel (a) depicts the standard monotonicity case, where type boundaries are horizontal and determined by fixed shares within each treatment type. Panel (b) shows the separable model, where the boundaries are smooth functions of U_T given by the MMR curves $\rho(1, U_T)$ and $\rho(0, U_T)$. The crossing of these curves produces local regions where the mediator response differs across adjacent treatment-type blocks.

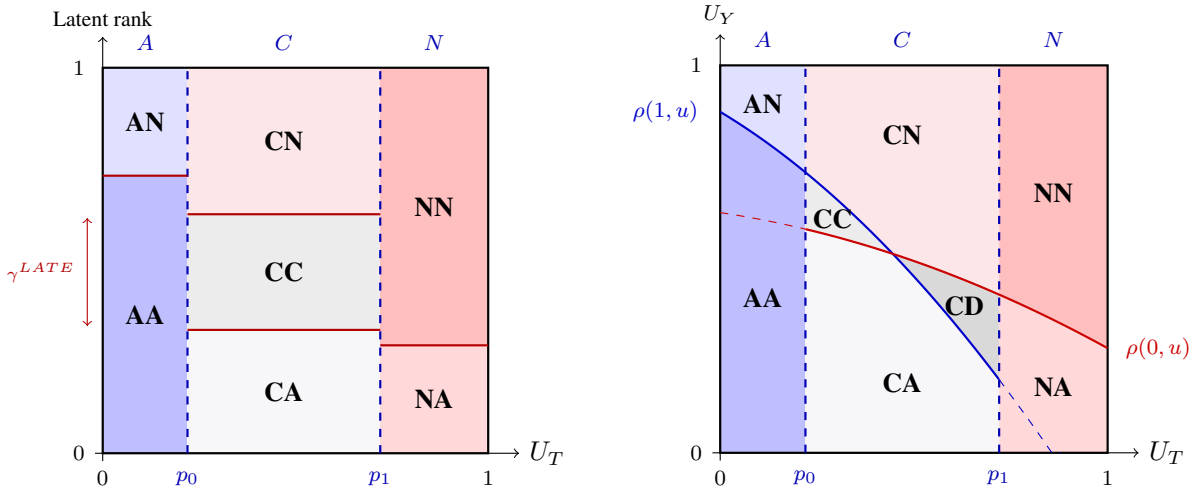
Table G.1 provides the corresponding joint-type classification. Each type is defined by its treatment and mediator responses to the instrument, and the table shows how mediator separability expands the type space relative to the monotonic benchmark.

Table G.1: Joint Types under Treatment Monotonicity and Mediator Separability

	T-Never-Taker (N)		T-Complier (C)				T-Always-Taker (A)	
	NN	NA	CN	CC	CA	CD	AN	AA
$(T(z_0),\ M(T(z_0)))$	$(0,0)$	$(0,1)$	$(0,0)$	$(0,0)$	$(0,1)$	$(0,1)$	$(1,0)$	$(1,1)$
$(T(z_1),\ M(T(z_1)))$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$	$(1,1)$	$(1,0)$	$(1,0)$	$(1,1)$

Notes: Treatment defiers ($T(z_0) > T(z_1)$) are ruled out by Assumption 3. Under mediator separability (54), the MMR curves $\rho(1,\cdot)$ and $\rho(0,\cdot)$ may cross, producing the CD (mediator defier) type alongside the standard CC (mediator complier) type. Each joint type is denoted by two letters: the first indicates treatment compliance (N = Never-taker, C = Complier, A = Always-taker) and the second indicates mediator compliance (N = Never-taker, C = Complier, A = Always-taker, D = Defier).

The key implication is that the two frameworks rely on different sources of variation. Standard mediator monotonicity restricts mediator responses across treatment states, while mediator separability restricts responses within treatment states. The conditional Wald ratio exploits within-treatment compositional variation across adjacent type blocks, which is captured by separability but not by standard monotonicity.



(a) Monotonicity, no separability

(b) Separability, no monotonicity

Figure G.1: Joint types under two mediator assumptions

Notes: Both panels display the (U_T, U_Y) unit square partitioned into treatment-type columns by p_0 and p_1. **Panel (a)**: Under mediator monotonicity ($M(1) \geq M(0)$) without separability, type boundaries are horizontal bands whose heights equal conditional type shares $\Pr(S_M \mid S_T)$; boundaries jump across columns. The CC band width equals γ^{LATE}. Seven types. **Panel (b)**: Under mediator separability ($M = \mathbf{1}[\rho(T, U_T) \geq U_Y]$) without mediator monotonicity, type boundaries are the smooth MMR curves $\rho(1, U_T)$ (blue) and $\rho(0, U_T)$ (red), drawn solid on their relevant domain and dashed outside it. The curves cross at $U_T \approx 0.45$ in the complier column: CC types (mediator compliers) arise where $\rho(1,u) > \rho(0,u)$, CD types (mediator defiers) where $\rho(0,u) > \rho(1,u)$. Eight types. Without MCA, type probabilities $\neq$ region areas; with MCA ($U_T \perp\!\!\!\perp U_Y$), probabilities $=$ areas.

Remark 12 (Assumption hierarchy and connection to LIV)**.** *The full mediator separability condition ($\rho(t,\cdot)$ globally decreasing) implies compositional mediator monotonicity (Assumption 4) but is strictly stronger: separability restricts ρ globally, while compositional monotonicity restricts it only on the adjacent blocks that generate the conditional first stage. Standard mediator monotonicity* (**??**) *is logically independent of both: it restricts ρ* across *treatment levels ($\rho(1,u) \geq \rho(0,u)$), while compositional monotonicity restricts ρ* within *each treatment level. In the LIV formulation, standard mediator mono-*

tonicity is testable because both MMR curves are identified by Corollary 2. This hierarchy determines each condition's role: MCA plus treatment monotonicity justify the conditional Wald construction; compositional monotonicity upgrades that ratio to a local mediator effect; standard mediator monotonicity is not required for Theorem 3.

G.3 Kitagawa-Style Falsification Tests: Implementation Details

Monotonicity conditions for the conditional test. The conditional Kitagawa test uses two ingredients: conditional exogeneity ($Y(t,m) \perp\!\!\!\perp S_T \mid T$, established in Lemma 1) and a no-defier monotonicity restriction for the mediator within each treatment stratum. Compositional mediator monotonicity (Assumption 4) provides the latter.

Assumption 5 (Orientation for the induced mediator first stage within strata)**.** *For each $t \in \{0,1\}$, orient (z_0, z_1) within the $T = t$ stratum so that*

$$\Delta_t \equiv \Pr(M = 1 \mid Z = z_1, T = t) - \Pr(M = 1 \mid Z = z_0, T = t) \geq 0.$$

This is a sign normalization, not a behavioral restriction. Compositional mediator monotonicity (Assumption 4) determines the sign of Δ_t and provides the "no defiers" condition required by Kitagawa (2015). Specifically:

- *Within $T = 1$:* Changing Z from z_0 to z_1 adds C-types alongside the incumbent A-types. Assumption 4 gives $M(1,A) \geq M(1,C)$ almost surely, so $\{M(1,C) = 1\} \subseteq \{M(1,A) = 1\}$ almost surely. This is the set inclusion that underlies the Kitagawa proof: in finite-unit language, no unit has $M = 1$ under the C-type propensity but $M = 0$ under the A-type propensity (no mediator defiers with respect to the A-vs-C margin).

- *Within $T = 0$:* Changing Z from z_0 to z_1 removes C-types, leaving only N-types. Assumption 4 gives $M(0,C) \geq M(0,N)$ almost surely, so $\{M(0,N) = 1\} \subseteq \{M(0,C) = 1\}$ almost surely. Removing C-types can only reduce the share with $M = 1$; in finite-unit language, this rules out mediator defiers with respect to the N-vs-C margin. The treatment-equation orientation gives $\Delta_0 \leq 0$, so the sign normalization swaps (z_0, z_1) within $T = 0$.

Compositional intuition. The conditional first stage $Z \Rightarrow M$ within $T = t$ is purely compositional. Consider the subpopulation with $T = 0$. When $Z = z_0$, the $T = 0$ stratum contains never-takers (N) and treatment compliers (C). When $Z = z_1$, treatment compliers satisfy $T(z_1) = 1$ and thus *exit* the $T = 0$ stratum, leaving only never-takers. Since $M(0,C) \geq M(0,N)$ almost surely (compositional monotonicity), the exiting compliers have weakly higher mediator rates, so their departure shifts mass from $M = 1$ to $M = 0$ within the stratum. An analogous argument holds in the $T = 1$ stratum, where the entry of treatment compliers (with weakly lower mediator rates than always-takers, by $M(1,A) \geq M(1,C)$ almost surely) shifts the mediator composition in the opposite direction.

Implementation via moment inequalities. Because the conditional Kitagawa inequalities (88)–(89) must hold for all Borel sets A, we test them over a rich finite class. A convenient choice is threshold sets

$A_j = (-\infty, y_j]$ for a grid $\{y_j\}_{j=1}^J$ (e.g. empirical quantiles of Y). Define the moments

$$g_{0j}^{(t)} \equiv \Pr(Y \le y_j,\, M = 0 \mid Z = z_0, T = t) - \Pr(Y \le y_j,\, M = 0 \mid Z = z_1, T = t),$$
$$g_{1j}^{(t)} \equiv \Pr(Y \le y_j,\, M = 1 \mid Z = z_1, T = t) - \Pr(Y \le y_j,\, M = 1 \mid Z = z_0, T = t),$$

so the null is $g_{mj}^{(t)} \ge 0$ for all (t, m, j).

Let $\widehat{g}_{mj}^{(t)}$ denote sample analogues (cell frequencies within $\{Z = z, T = t\}$), and let $\widehat{\sigma}_{mj}^{(t)}$ be corresponding standard errors. A standard "max violation" statistic is

$$\widehat{T}_{\mathrm{MCA}} \;=\; \max_{t\in\{0,1\},\, m\in\{0,1\},\, j\le J} \left\{ -\frac{\sqrt{n}\,\widehat{g}_{mj}^{(t)}}{\widehat{\sigma}_{mj}^{(t)}} \right\}_+ .$$

Critical values can be obtained via a nonparametric bootstrap (individual- or cluster-level, matching the sampling design), or via modern moment-inequality methods applied to the stacked vector of inequalities.

Testing without conditional monotonicity. Without conditional monotonicity for the induced first stage within $T = t$, one can replace (88)–(89) with instrumental-inequality tests that do not impose monotonicity for the binary conditional-IV model of $(Y, M, Z) \mid (T = t)$. This alternative remains a falsification test for the conditional-IV implications implied by mediated confounding, but it does not deliver the same sharp Kitagawa inequality characterization.

H Post-Double Selection for High-Dimensional Controls

This appendix extends the linear identification of Section 4 to settings with a high-dimensional vector of candidate controls. The mediation system contains three structural equations rather than one, so the standard post-double selection procedure of Belloni et al. (2014) (henceforth BCH) requires a fourth nuisance function to handle the instrument's dependence on covariates. We refer to the resulting procedure as *quadruple-LASSO*: four parallel LASSO regressions—one each for Y, T, M, and Z—whose union of selected variables forms the control set for the conditional 2SLS of Section 4.2.

H.1 Setup

Let $X_i \in \mathbb{R}^{d_x}$ denote the user-supplied controls and let $W_i = \phi(X_i) \in \mathbb{R}^p$ denote a rich expansion of X_i (powers, interactions, basis transformations); p may be large relative to n, including $p > n$. The structural model with controls augments (21)–(23) by allowing each variable to depend nonparametrically on X:

$$T_i = g_T(X_i) + \pi_1 Z_i + \varepsilon_{i,T}, \tag{156}$$

$$M_i = g_M(X_i) + \gamma T_i + \varepsilon_{i,M}, \tag{157}$$

$$Y_i = g_Y(X_i) + \tau T_i + \theta M_i + \varepsilon_{i,Y}, \tag{158}$$

together with the *fourth* nuisance function

$$g_Z(X_i) \equiv \mathrm{E}[Z_i \mid X_i], \tag{159}$$

which is identically zero under exogenous instrument assignment but generally nontrivial in quasi-experimental settings. Identifying assumptions remain the IV exogeneity (24) and MCA (17), augmented for the high-dimensional appendix by the strengthening

$$\mathrm{E}[\varepsilon_{i,T} \mid X_i] = \mathrm{E}[\varepsilon_{i,M} \mid X_i] = \mathrm{E}[\varepsilon_{i,Y} \mid X_i] = 0, \tag{160}$$

which is stronger than the unconditional moment conditions used in Section 4 and is needed for valid residualization against X.

Conditional means versus structural components. The four LASSO regressions in Algorithm 1 below regress $V \in \{Y, T, M, Z\}$ on the rich expansion W and therefore estimate the *conditional means*

$$m_V(X_i) \equiv \mathrm{E}[V_i \mid X_i], \quad V \in \{Y, T, M, Z\}, \tag{161}$$

not the structural components g_V. Under the zero-conditional-mean errors (160) and the structural equations (156)–(158),

$$m_T = g_T + \pi_1 g_Z, \quad m_M = g_M + \gamma m_T, \quad m_Y = g_Y + \tau m_T + \theta m_M, \quad m_Z = g_Z. \tag{162}$$

The high-dimensional appendix is therefore stated in terms of the four conditional means m_V, not the structural functions g_V: residualization, sparsity, and Neyman orthogonality must all reference m_V,

since these are the objects the LASSO actually targets.

The post-selection procedure relies on each conditional-mean nuisance function being well approximated by a sparse linear combination of the rich expansion.

Assumption 6 (Approximate Sparsity of Conditional Means)**.** *For each $V \in \{Y, T, M, Z\}$, there exists a coefficient vector $\pi_V^* \in \mathbb{R}^p$ with at most s nonzero components and an approximation error r_V such that $m_V(X_i) = W_i'\pi_V^* + r_V(X_i)$, with $n^{-1}\sum_i r_V(X_i)^2 = O(s/n)$ and $s \log p = o(\sqrt{n})$.*

This is the standard BCH condition: s may grow with n but slower than $\sqrt{n}/\log p$.

H.2 The Quadruple-LASSO Algorithm

Identification of (θ, τ, γ) in the high-dimensional setting follows from partialling each variable on its conditional-mean nuisance function. Define

$$\tilde{V}_i \equiv V_i - m_V(X_i), \quad V \in \{Y, T, M, Z\}. \tag{163}$$

Substituting (156)–(158) together with the relations (162) yields the partialled system

$$\tilde{T}_i = \pi_1 \tilde{Z}_i + \varepsilon_{i,T}, \quad \tilde{M}_i = \gamma \tilde{T}_i + \varepsilon_{i,M}, \quad \tilde{Y}_i = \tau \tilde{T}_i + \theta \tilde{M}_i + \varepsilon_{i,Y}, \tag{164}$$

which has the same structural form as (21)–(23) and is identified by Theorem 1. The partialled equations hold pointwise only when residualization is by conditional means m_V, not by structural components g_V: residualizing by g_V leaves leftover terms $\pi_1 g_Z$ in $\tilde{T}$, γg_T in $\tilde{M}$, and $\tau g_T + \theta g_M$ in $\tilde{Y}$ that contaminate the partialled system.

In a finite sample, each m_V is unknown and must be estimated. The quadruple-LASSO procedure does so by running four parallel LASSO regressions on W_i, then computing the structural parameters by 2SLS on the union of selected controls.

Algorithm 1 (Quadruple-LASSO Post-Double Selection)**.**

1. ***Construct the rich expansion*** $W_i = \phi(X_i)$ *comprising the user-supplied controls together with their squares, cubes, and pairwise interactions; pure indicators are included without powers (which would be mechanically equal to the indicator).*
2. ***Absorb high-cardinality fixed effects*** *by within-group demeaning of $(Y_i, T_i, M_i, Z_i, W_i)$.*
3. ***Run four LASSOs*** *on the demeaned variables: for each $V \in \{Y, T, M, Z\}$,*

$$\hat{\beta}_V = \arg\min_{\beta} \tfrac{1}{2n_w} \textstyle\sum_i w_i \, (V_i - W_i'\beta)^2 + \lambda_V \sum_j \psi_{V,j} |\beta_j|, \tag{165}$$

   with $n_w = \sum_i w_i$ and the penalty levels λ_V and loadings $\psi_{V,j}$ computed from the iterative plug-in formulas of Belloni et al. (2014) and Belloni et al. (2012), adapted to heteroskedasticity-, cluster-, and weight-robust inference following Belloni et al. (2018); under clustering, the penalty level uses $\sqrt{G_{\text{eff}}}$ rather than $\sqrt{n}$.
4. ***Form the union selection set***

$$S = S_Y \cup S_T \cup S_M \cup S_Z \cup X_{\text{keep}}, \qquad S_V = \{j : \hat{\beta}_{V,j} \neq 0\}, \tag{166}$$

where $X_{keep} \subseteq X$ is an optional subset of controls flagged by the researcher to be retained unconditionally and omitted from the LASSO penalty (e.g., when domain knowledge mandates inclusion regardless of selection).

5. ***Run the conditional 2SLS*** *of Section 4.2 with $W_i^{(S)}$ as controls, where $W_i^{(S)}$ contains only the components of W_i indexed by S. Compute $\hat{\theta}, \hat{\tau}, \hat{\gamma}$, the first-stage F-statistics of Remark 2, and the κ-sensitivity diagnostics of Section 5.2.1 on the post-selection design.*

A natural-looking variant would regress M_i on W_i *and* T_i in step 3 of Algorithm 1. This is incorrect: doing so allows the procedure to absorb into T predictive power that should be allocated to X, dropping from S_M controls whose effect on M operates through their correlation with T. The mediator-selection regression therefore uses X alone; T enters only the structural 2SLS in step 5.

H.3 Asymptotic Normality

Proposition 10 (Post-Selection Asymptotic Normality)**.** *Suppose the IV exogeneity (24), MCA (17), the zero-conditional-mean errors (160), Assumption 6 (approximate sparsity of conditional means), and the standard BCH regularity conditions on the design matrix W and the score moments hold. Suppose further that the residualized mediation Jacobian*

$$G \equiv -\,\mathrm{E}\begin{pmatrix} \tilde{Z}\,\tilde{T} & \tilde{Z}\,\tilde{M} & 0 \\ \tilde{T}^2 & \tilde{T}\,\tilde{M} & 0 \\ 0 & 0 & \tilde{Z}\,\tilde{T} \end{pmatrix} \tag{167}$$

is nonsingular with singular values bounded away from zero, where $\tilde{V} \equiv V - m_V(X)$ as in (163). Apply Algorithm 1 with the iterative plug-in penalty. Then the post-selection 2SLS estimator $(\hat{\theta}, \hat{\tau}, \hat{\gamma})$ is $\sqrt{n}$-consistent and asymptotically normal:

$$\sqrt{n}\big[(\hat{\theta}, \hat{\tau}, \hat{\gamma}) - (\theta, \tau, \gamma)\big]' \xrightarrow{d} \mathcal{N}(0, \Omega), \tag{168}$$

with $\Omega = G^{-1}\Sigma G^{-1\prime}$ for $\Sigma = \mathrm{E}[\psi_i \psi_i']$ with ψ_i the orthogonal moment vector defined in Step 1 of the proof. Ω is consistently estimated by the standard sandwich formula on the post-selection design, using heteroskedasticity- or cluster-robust components.

Proof sketch. The result is a direct application of Belloni et al. (2014, Theorem 1) extended to the three-equation mediation system. The argument proceeds in three steps.

Step 1: Neyman orthogonality at the four conditional-mean nuisance functions. The mediation moment conditions, after partialling by m_Y, m_T, m_M, m_Z as in (163), are

$$\mathrm{E}[(\tilde{Y} - \tau\tilde{T} - \theta\tilde{M})\tilde{Z}] = 0, \quad \mathrm{E}[(\tilde{Y} - \tau\tilde{T} - \theta\tilde{M})\tilde{T}] = 0, \quad \mathrm{E}[(\tilde{M} - \gamma\tilde{T})\tilde{Z}] = 0. \tag{169}$$

The first identifies θ, the second τ (using MCA to make $\tilde{T}$ exogenous in the outcome equation), and the third γ. Stack these into the score $\psi_i(\alpha, \eta)$ with $\alpha = (\tau, \theta, \gamma)'$ and $\eta = (m_Y, m_T, m_M, m_Z)$. The Gateaux derivative of $\mathrm{E}[\psi_i(\alpha, \eta)]$ with respect to a perturbation h_V in m_V at the truth equals zero: under (160), the residual $\tilde{Y} - \tau\tilde{T} - \theta\tilde{M} = \varepsilon_Y$ satisfies $\mathrm{E}[\varepsilon_Y \mid X] = 0$, and analogously $\mathrm{E}[\tilde{T} \mid X] = \mathrm{E}[\tilde{Z} \mid X] = 0$, so each first-order perturbation term has mean zero by iterated expectations.

This is the Neyman-orthogonality property required for post-selection inference, and it holds only when residualization uses m_V, not g_V: the structural components carry residual conditional dependence on X via $\beta_{TZ} g_Z$, γg_T, etc., that breaks $\mathrm{E}[\tilde{V} \mid X] = 0$.

Step 2: Sufficient nuisance accuracy from the union-selection set. Under Assumption 6 and the iterative plug-in penalty, Belloni et al. (2014, Theorem 1) and Belloni et al. (2012, Theorem 1) establish prediction-norm convergence rates

$$\|\hat{m}_V - m_V\|_{P,2} = O_p\big(\sqrt{s \log p / n}\big) = o_p(n^{-1/4}), \quad V \in \{Y, T, M, Z\}, \tag{170}$$

on the union-selected control set S. The argument does not require exact support recovery of any π_V^*: small but nonzero coefficients may be omitted by the LASSO without invalidating the prediction-norm rate, and the orthogonal score in Step 1 absorbs any first-order effect of nuisance estimation error.

Step 3: Apply the BCH normality result. With Steps 1–2 established, the post-selection 2SLS falls within the BCH framework. The orthogonal stacked-moment expansion gives

$$\sqrt{n}(\hat{\alpha} - \alpha) = -G^{-1}\, n^{-1/2} \sum_{i=1}^{n} \psi_i(\alpha, \eta_0) + o_p(1),$$

where the rank condition on (167) ensures G^{-1} exists. A central limit theorem applied to the leading term delivers $\mathcal{N}(0, G^{-1}\Sigma G^{-1\prime})$, with $\Sigma = \mathrm{E}[\psi_i \psi_i']$. The sandwich variance estimator on the post-selection design is consistent under the maintained moment and empirical-process conditions (Belloni et al., 2018). □

H.4 Compatibility with the Paper's Diagnostics

First-stage F-statistics. The two first-stage diagnostics of Remark 2, $F_{T,Z|X}$ and $F_{M,Z|T,X}$, are computed post-selection on the design that includes $W^{(S)}$ as controls. By Proposition 10 and the prediction-norm rate in Step 2 of its proof, the population residualization against m_T, m_M, m_Z is asymptotically equivalent to the sample residualization against $W^{(S)}$. The post-selection F-statistics therefore converge in distribution to their oracle counterparts, and the Olea–Pflueger critical values $c(K, \tau, \alpha)$ apply on the post-selection design without modification. The closed-form decomposition $F_{M,Z|T,X} \overset{\text{iid}}{=} (n-3)\, r_{TZ}^2 \rho_{TM}^2 / (1 - \rho_{TM}^2)$ from Remark 2 continues to hold, with the relevant correlations now defined on the post-selection partialled variables. Both weak-identification channels—weak instrument ($\pi_1 \to 0$) and weak collider ($\rho_{TM} \to 0$)—remain visible in the diagnostic.

κ-based sensitivity. The breakeven $\kappa^* = \theta \cdot \mathrm{cov}(\varepsilon_T, \varepsilon_M)$ defined in Section 5.2.1 is a population object that does not depend on which sample-level controls are used to estimate residuals. Under Assumption 6, the zero-conditional-mean errors (160), and the prediction-norm rate (170) in Step 2 of the proof of Proposition 10, the post-selection residuals $\hat{\varepsilon}_T$ and $\hat{\varepsilon}_M$ converge to the structural residuals at the same rate as the structural coefficients. The sample covariance $\widehat{\mathrm{cov}}(\varepsilon_T, \varepsilon_M)$ remains consistent for the population covariance, so the breakeven $\hat{\kappa}^*$, the inferential breakeven $\hat{\kappa}^*_{\text{CI}}$, and the correlation-scale breakeven $\hat{\rho}^*_{TY}$ retain their interpretation post-selection without modification. The same applies to the κ-adjusted estimator $\hat{\theta}_{\text{adj}}(\kappa_0)$ and its confidence interval.

H.5 Practical Notes

Remark 13 (Cross-fitting and double machine learning)**.** *The procedure above is the post-double selection version: nuisance functions are estimated and structural parameters computed on the same sample. Cross-fitting (or double/debiased machine learning, Chernozhukov et al., 2018) splits the sample into folds, estimates each nuisance function on the complement of a fold, and averages the structural estimates across folds. Cross-fitting buys additional protection against overfitting bias at the cost of computational overhead. The four-equation framework above is fully compatible with cross-fitting: each fold complement supplies an estimate of all four nuisance functions, and the structural moments* (169) *are evaluated on the held-out fold.*

Remark 14 (When to use post-double selection)**.** *The procedure is most useful when (i) the researcher faces many candidate covariates without strong prior guidance on which to include, (ii) functional-form misspecification is a concern that the rich expansion can address, and (iii) the sample size is large enough relative to p for approximate sparsity to be credible. With a small set of well-motivated controls, manual specification is preferable.*

Remark 15 (Replication)**.** *For replicability, applied work should report the four selection sets S_Y, S_T, S_M, S_Z, the four chosen penalty values $\lambda_Y, \lambda_T, \lambda_M, \lambda_Z$, and the dimension p of the rich expansion alongside the standard regression output. The reference implementation* `iv_mediation_applied.py` *(see footnote in Section 4.2) provides these objects in the structured return value of the analysis function and supports the iterative plug-in penalty under heteroskedasticity, clustering, and weights.*

I Empirical Application Sources

Table I.1: Overview of empirical applications

Study	Application and variables
Becker & Woessmann (2009, QJE)	Revisit Weber's claim that Protestantism increased economic prosperity through a Protestant work ethic. Using county-level data from late-nineteenth-century Prussia, they instead argue that Protestantism raised prosperity primarily by increasing literacy and human capital, since Protestant doctrine emphasized individual Bible reading. *Variables:* Y is economic prosperity in Prussian counties; T is the Protestant share; M is the literacy rate; Z is distance to Wittenberg.
Autor et al. (2020, AER)	Study whether rising import competition from China contributed to ideological realignment and political polarization in the United States. Using local labor markets' exposure to Chinese import competition from 2000 to 2008, they find that trade-exposed areas shifted toward Republican presidential candidates, with broader evidence of both rightward shifts and increased ideological polarization. *Variables:* Y is the change in the Republican two-party presidential vote share from 2000 to 2008; T is the change in U.S. import exposure from China per capita from 2000 to 2008; M is the change in manufacturing employment share from 2000 to 2007 from Autor, Dorn, and Hanson (2013); Z is the other-high-income-country Bartik instrument for Chinese import exposure.
Dippel et al. (2022, EJ)	Study whether trade exposure contributed to rising support for nationalist and right-populist parties in Germany. Using regional variation in exposure to imports from low-wage countries, they find that import competition increased far-right vote shares, while export exposure had the opposite effect, and identify trade-induced labor market adjustment as an important mechanism. *Variables:* Y is the change in far-right vote share; T is net import exposure per worker at the Kreis level; M is the log change in employment in trade-dependent manufacturing sectors; Z is the other-country net-supply instrument.
Devoto et al. (2012, AEJ:EP)	Study households' willingness to pay for private water connections when the connection can be financed through credit. They find that private water connections improved household well-being not primarily through health or income gains, but by increasing time available for leisure and reducing inter- and intra-household conflicts over water. *Variables:* Y is an indicator for whether the household reports that life improved after 24 months; T is take-up of the private water connection; M is the leisure and social well-being index; Z is assignment to the treatment group.

Continued on next page

Table I.1: Overview of empirical applications *(continued)*

Study	Application and variables
Arnsbarger et al. (2026, wp)	Study whether women's labor market participation enabled political mobilization after the U.S. Civil War. Linking Union Army enlistments to the 1860 and 1870 censuses, they show that the wives and daughters of disabled Union Army veterans were more likely to enter the labor force, and that towns with both higher female labor force participation and more disabled veterans saw greater Temperance Crusade activity in 1873. *Variables:* Y is an indicator for Temperance Crusade protest activity in 1873; T is the standardized share of disabled or wounded Union Army veterans; M is standardized female labor force participation in 1870; Z is the share of soldiers excluded from combat.
Froelich & Huber (2017, JRSS-B) BHPS example	Use data from the British Household Panel Survey to study whether education improves social functioning, and whether this effect operates through income. They find that education increases social functioning, but their estimates suggest that this effect operates mostly through channels other than annual income. *Variables:* Y is the social functioning index; T is an indicator for having more than lower secondary education; M is annual individual income; Z_1 is exposure to the 1971 increase in the UK minimum school leaving age; Z_2 is windfall income from accident claims, redundancy payments, lottery wins, and other lump-sum payments.
Froelich & Huber (2017, JRSS-B) Job Corps example	Use data from the U.S. Job Corps experiment to study whether enrollment in the program increased earnings among young women from low-income households, and whether this effect operated through increased labor supply. Their estimates suggest that Job Corps increased weekly earnings primarily by increasing hours worked rather than by raising hourly wages. *Variables:* Y is weekly earnings in the third year after randomization; T is enrollment in Job Corps in the first or second year after randomization; M is hours worked per week in the third year after randomization; Z_1 is randomized treatment assignment; Z_2 consists of indicators for the number of children in the household younger than 6 and younger than 15.