Identifying Causal Mediation Effects with a Single Instrument*

Andreas Ferrara† Robert Gold‡ Stephan Heblich§ Rodrigo Pinto¶

May 22, 2026

Abstract

Instrumental variables (IV) methods identify causal treatment effects in the presence of endogeneity but do not reveal how these effects operate. With a single instrument, the total effect is identified, but it cannot be decomposed into direct and indirect components through a mediator. This paper shows that such a decomposition is feasible under *mediated confounding*, which restricts how unobservables link treatment and outcome while allowing the mediator to remain endogenous. Under this condition, the same instrument identifies both direct and indirect (mediated) effects. We establish identification in a linear model and show that estimation can be implemented using standard two-stage least squares. We also provide diagnostic tests and sensitivity measures for the identifying assumption and illustrate the approach in simulations and applications.

Keywords: Instrumental Variables, Mediation Analysis, Causal Mechanisms, Identification, Two-Stage Least Squares

JEL Codes: C26, C36

*Preliminary and incomplete. This draft is a work in progress and should not be cited or circulated.

†University of Pittsburgh, Department of Economics, and NBER. Email: a.ferrara@pitt.edu.

‡IfW Kiel and CESifo. Email: Robert.Gold@ifw-kiel.de.

§University of Toronto, Department of Economics, and NBER. Email: stephan.heblich@utoronto.ca.

¶University of South Florida. Email: rodrigopinto@usf.edu

1 Introduction

In evaluating the impact of a treatment on an outcome, applied economists often want to move beyond the total effect and understand *how* it operates. Does a job training program improve earnings by increasing educational attainment? Do improved neighborhoods affect children's outcomes by reducing exposure to poverty? Does education improve health through behavior or through income? In many such settings, researchers have a credible instrument for the treatment T and can estimate its effect on both an intermediate variable M and a final outcome Y. Yet standard instrumental variable methods cannot quantify what share of the total effect on Y operates *through* the mediator M. The challenge is not only technical: policy often targets mechanisms rather than treatments, so identifying which channels matter is central to implementing effective evaluations.

The main difficulty in moving beyond the total effects stems from a fundamental identification problem. With a single instrument that shifts the treatment but does not directly shift the mediator, all instrument-induced variation in M operates through T. As a result, we never observe variation in the mediator holding the treatment fixed. However, this is exactly the variation needed to separate the direct effect of T on Y from the indirect effect through M. Without it, many different decompositions of the total effect are consistent with the data. This intuition is formalized as a non-identification result in Section 2.

Existing approaches resolve this problem by adding either instruments or assumptions. The two-instrument method of Frölich and Huber (2017b) and Rudolph et al. (2024) relies on a second instrument that shifts the mediator independently of the treatment. While conceptually straightforward, this requirement is often difficult to satisfy in practice. Moreover, when different instruments identify different local average treatment effects, the resulting decomposition may correspond to different complier populations, complicating interpretation. Sequential ignorability (Imai et al., 2010, 2013) instead assumes that the mediator is as good as randomly assigned conditional on treatment and covariates. This effectively rules out unobserved confounding between M and Y, which is rarely credible when the mediator is itself an outcome influenced by unobservables that also affect Y. These approaches therefore illustrate a common trade-off: either require additional instruments, or impose strong exogeneity assumptions on the mediator. Our approach takes a different route. Rather than requiring a second instrument or assuming that the mediator is exogenous, we maintain endogeneity of both the treatment and the mediator, but impose structure on how unobservables link treatment and outcome. This shifts the identifying assumption away from the mediator-outcome relationship and onto the nature of confounding itself.

This paper shows that mediation analysis with a single instrument becomes feasible once one imposes such structure on the *source* of confounding. We introduce the condition of *mediated confounding*: the unobserved determinants of the treatment are independent of the unobserved determinants of the outcome, while the mediator may remain correlated with both. The key point is that both T and M remain endogenous. The assumption does not eliminate endogeneity; instead, it restricts how unobservables link treatment and outcome, requiring that any such link operates entirely through the mediator. In this sense, the assumption targets exactly the source of the identification problem described above.

The identifying mechanism has a simple intuition. Under mediated confounding, the instrument can be used twice. First, it shifts the treatment in the usual way. Second, once we condition on the treatment, it generates variation in the mediator that is independent of the outcome. The reason is that conditioning

on T induces a statistical relationship between the instrument and the unobserved determinants of the mediator, while the restriction on confounding ensures that the instrument remains exogenous for the outcome. This creates a valid source of variation for identifying the effect of the mediator. We formalize this argument and show that, under mediated confounding, the instrument is both exogenous for the outcome conditional on the treatment (Proposition 1) and relevant for the mediator conditional on the treatment.

We formalize this argument in a linear framework and show that all components of the mediation decomposition are identified using a single instrument. Under mediated confounding, the instrument is exogenous for the outcome conditional on the treatment and relevant for the mediator conditional on the treatment, allowing us to recover both the direct effect of T on Y and the effect of M on Y. The resulting decomposition is internally consistent: the sum of direct and indirect effects coincides with the standard Wald ratio (Corollary 1). Estimation can be implemented using familiar two-stage least squares procedures. In practice, this amounts to instrumenting the mediator using the original instrument while controlling for the treatment, so that the approach can be applied using standard econometric software.

We extend the analysis beyond linearity using the local instrumental variable approach of Heckman and Vytlacil (1999). In this setting, mediation effects are identified from changes in conditional expectations with respect to the propensity score. A key distinction from standard treatment effect analysis is that mediation operates through two stages: the instrument first shifts treatment participation and then alters the distribution of the mediator within treatment strata. As a result, mediation effects are identified from curvature in these conditional expectations rather than from their slope (Theorem 2).

When the instrument is binary, this logic simplifies. The conditional Wald ratio identifies a local mediator effect within each treatment stratum under mediated confounding, treatment monotonicity, and a compositional mediator monotonicity condition that is weaker than the standard requirement $M(1) \geq M(0)$ (Theorem 3). Because both direct and indirect effects are identified from the same source of variation, they correspond to the same complier population.

The identifying assumption is stronger than standard IV conditions and must be assessed carefully in applications. We provide two complementary tools for this purpose. First, when additional instruments are available, we develop specification tests that directly evaluate whether the data are consistent with mediated confounding. These include tests of overall model fit as well as tests that isolate the moment condition implied by the assumption. Second, in the baseline case with a single instrument, we develop a sensitivity analysis that quantifies how large a violation of the assumption would need to be to overturn the estimated mediator effect. This approach adapts the omitted variable bias framework of Cinelli and Hazlett (2024) to the conditional IV setting, providing a transparent and interpretable measure of robustness.

These tools are particularly useful in the common empirical setting where a researcher has a credible instrument for the treatment but no separate instrument for the mediator. In such cases, standard IV methods identify total effects but provide little guidance on mechanisms. Our framework allows researchers to assess whether a proposed mediator can account for the treatment effect and to quantify the importance of alternative channels when it cannot.

To evaluate the performance of the approach, we conduct simulation exercises similar to the design of Frölich and Huber (2017b). We compare the single-instrument estimator directly to their two-instrument benchmark. The simulations show that the proposed method recovers comparable estimates

for both direct and indirect effects. At the same time, they highlight an important practical consideration: because identification relies on conditional variation in the mediator, the conditional first stage can be substantially weaker than the standard first stage. This underscores the importance of reporting conditional first-stage diagnostics and using robust inference when necessary.

We also illustrate the empirical usefulness of the approach by applying it to seven existing IV settings from economic history, political economy, development, labor, and program evaluation. The applications revisit studies in which researchers have a credible instrument for a treatment and a substantively important candidate mediator, but typically no separate instrument for the mediator. Across these settings, the estimator recovers the original total IV effects and decomposes them into direct and indirect components using the same source of variation. The results show that the method can summarize mechanism evidence in a transparent way, while the conditional first-stage and sensitivity diagnostics clarify when the mediated channel is strongly supported by the data and when it should be interpreted more cautiously.

This paper contributes to a growing literature on mediation analysis in instrumental variable settings. Existing approaches differ in the margin on which they impose identifying structure. The two-instrument approach, in which separate instruments shift the treatment and the mediator, is developed by Frölich and Huber (2017b) and extended by Rudolph et al. (2024). Several contributions study settings with an exogenous treatment and an instrument for the mediator, including Robins and Greenland (1992), Mattei and Mealli (2011), Imai et al. (2013), and Attanasio et al. (2020). Other work considers an endogenous treatment with an instrument for the mediator under parametric assumptions (Dunn and Bentall, 2007; Albert, 2008; Small, 2012; Chen et al., 2019; Joffe et al., 2008). In the education–health mediation literature, Brunello, Fort, Schneeweis and Winter-Ebmer (2016) decompose the effect of education on health into channels through health behaviors. Their decomposition relies on a selection-on-observables strategy combined with aggregation and gender-differencing; their instrumental-variables analysis, by contrast, estimates only the total effect, since the absence of a credible instrument for health behaviors prevents IV from decomposing the gradient. Our MCA framework addresses precisely this gap—it permits a mediation decomposition with a single instrument for the treatment alone, without requiring a second instrument for the mediator. Heckman et al. (2013) decompose long-run treatment effects in a randomized intervention into channels operating through latent skill factors identified by multiple imperfect measurements—a methodologically distinct approach that applies to a different data setting from ours.

Our approach differs in how it resolves the identification problem. Rather than introducing additional instruments or assuming that the mediator is exogenous, we maintain endogeneity of both the treatment and the mediator but impose structure on how unobservables link treatment and outcome. In particular, while sequential ignorability restricts the mediator-outcome relationship, and multi-instrument approaches introduce additional sources of variation, our framework instead restricts the treatment-outcome link, leaving the mediator-outcome relationship unrestricted. In this sense, our approach complements existing methods by trading off additional identifying structure for weaker data requirements, while preserving a setting that is close to standard IV applications.

More broadly, our framework relates to a large applied literature that studies mechanisms by examining intermediate outcomes. While such analyses are often informal, our approach provides a formal framework for assessing whether a proposed mediator can account for the observed treatment effect and

for quantifying the contribution of alternative channels. In contrast to work that focuses on testing or bounding the role of mechanisms, we directly identify mediation effects under a single instrument by imposing structure on the source of confounding.

The remainder of the paper proceeds as follows. Section 2 presents the IV framework and the identification challenge. Section 3 introduces the mediated confounding assumption. Section 4 develops identification and estimation in the linear model. Section 5 presents specification tests and sensitivity analysis. Section 6 extends the framework beyond linearity. Section 7 presents simulation evidence. Section 8 discusses empirical applications. Section 9 concludes.

2 The Identification Problem in IV Mediation

We begin by presenting the standard IV framework and then formalizing the identification problem that arises when a mediator is introduced. The key point is that, with a single instrument, the total effect of the treatment is identified, but its decomposition into direct and indirect components is not.

Regularity assumptions. All observables and potential outcomes are random variables defined on a common probability space whose joint distribution admits regular conditional distributions with respect to every conditioning variable used in the paper. Conditional moments exist and are finite whenever invoked. Equalities, inequalities, and independence relations between conditional distributions are understood to hold almost surely with respect to the marginal distribution of the conditioning variables.

Notational convention for conditional dependence. For any random variables V, W, and U, we write $V \not\perp\!\!\!\perp W \mid U$ to mean that the conditional distribution of W given (V, U) is a non-trivial function of V on a positive-probability subset of U: there exist values $v \neq v'$ of V such that

$$\Pr(W \in A \mid V = v, U) \;\neq\; \Pr(W \in A \mid V = v', U) \quad \text{for some Borel event } A,$$

on a set of U-values with positive measure. This formulation covers categorical, continuous, and mixed-type variables. Under linearity with finite second moments, it is implied by—and equivalent to—the covariance form $\mathrm{Cov}(V, W \mid U) \neq 0$. The relation $V \perp\!\!\!\perp W \mid U$ is read analogously: the conditional distribution of W given (V, U) does not vary with V, almost surely in U.

2.1 Baseline IV Framework

We consider an endogenous treatment $T \in \mathcal{T}$, an outcome $Y \in \mathbb{R}$, an instrument $Z \in \mathcal{Z}$, and a vector of observed covariates X. Let $Y(t)$ denote the potential outcome under treatment level t, $T(z)$ the potential treatment under instrument value z, and $Y(t, z)$ the potential outcome under the joint intervention $(T, Z) = (t, z)$. The *core instrumental-variables assumptions* require the instrument to satisfy three conditions:

$$\textbf{Exclusion Restriction:} \quad Y(t, z) = Y(t) \quad \text{for all } (t, z) \in \mathcal{T} \times \mathcal{Z}, \tag{1}$$

$$\textbf{Exogeneity:} \quad Z \perp\!\!\!\perp \Big(\{Y(t)\}_{t \in \mathcal{T}},\, \{T(z)\}_{z \in \mathcal{Z}}\Big) \,\Big|\, X, \tag{2}$$

$$\textbf{Relevance:} \quad Z \not\perp\!\!\!\perp T \,\big|\, X. \tag{3}$$

Substantively, exclusion (1) requires that Z affects Y only through T; exogeneity (2) requires that, conditional on X, the instrument is as good as random; and relevance (3) requires that Z generates variation in T after partialling out X. Under (1)–(3), the Wald estimand

$$\frac{\text{Cov}(Y, Z \mid X)}{\text{Cov}(T, Z \mid X)}$$

is well-defined. Its causal interpretation requires one additional identifying assumption: linearity of the structural equations identifies the average treatment effect; monotonicity of the treatment response in Z identifies the local average treatment effect; separability of the selection equation identifies the marginal treatment effect. The choice among these assumptions determines the causal parameter but does not affect the mediation identification problem studied below.

2.2 Introducing the Mediator and the Decomposition

We now introduce a mediator $M \in \mathcal{M}$, a variable caused by treatment T that in turn causes outcome Y. Let $M(t)$ denote the potential mediator under treatment level t, and $Y(t, m)$ denote the potential outcome under treatment t and mediator m. The total effect of the treatment can be decomposed into a *direct effect*—the component operating through channels other than M—and an *indirect effect*—the component operating through the causal chain $T \to M \to Y$. Following Robins and Greenland (1992) and Imai et al. (2010), define the natural direct and indirect effects at reference treatment level t:

$$\text{NDE}(t) \equiv \text{E}\left[Y(1, M(t)) - Y(0, M(t))\right], \tag{4}$$

$$\text{NIE}(t) \equiv \text{E}\left[Y(t, M(1)) - Y(t, M(0))\right]. \tag{5}$$

The total effect admits two equivalent decompositions: $\text{TE} = \text{NDE}(0) + \text{NIE}(1) = \text{NDE}(1) + \text{NIE}(0)$. The two decompositions coincide when the direct effect does not depend on the mediator's potential value—that is, when there is no treatment–mediator interaction. Under the linear model developed below, this decomposition simplifies: the direct effect corresponds to the coefficient on the treatment, and the indirect effect corresponds to the product of the treatment-mediator effect and the mediator-outcome effect.

2.3 IV with a Mediator: What Is Identified

The *extended core IV assumptions* for the mediation setting are:

$$\textbf{Mediation Exclusion:} \quad M(t, z) = M(t) \;\text{ and }\; Y(t, m, z) = Y(t, m) \quad \text{for all } (t, m, z), \tag{6}$$

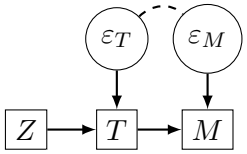
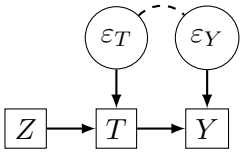
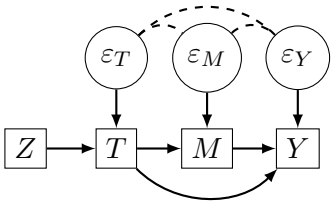
$$\textbf{Mediation Exogeneity:} \quad Z \perp\!\!\!\perp \Big(\{Y(t, m)\}_{(t,m)\in\mathcal{T}\times\mathcal{M}},\, \{M(t)\}_{t\in\mathcal{T}},\, \{T(z)\}_{z\in\mathcal{Z}}\Big) \,\Big|\, X, \tag{7}$$

$$\textbf{Mediation Relevance:} \quad Z \not\perp\!\!\!\perp T \,\big|\, X. \tag{8}$$

Mediation exclusion strengthens the standard exclusion restriction by requiring that the instrument affects neither the mediator nor the outcome directly: all effects operate through the treatment. Mediation exogeneity extends the independence condition to all potential outcomes, including those involving the mediator.

Table 1: The Identification Problem of Mediation Analysis with IV

A. Graphical Representation

Model I: *Intermediate Outcome M*	*Model II:* *Final Outcome Y*	*Model III:* *Structural Mediation Model*
		

B. Model Equations

$T = f_T(Z, \varepsilon_T)$	$T = f_T(Z, \varepsilon_T)$	$T = f_T(Z, \varepsilon_T),\ M = f_M(T, \varepsilon_M)$
$M = f_M(T, \varepsilon_M)$	$Y = f_Y(T, \varepsilon_Y)$	$Y = f_Y(T, M, \varepsilon_Y)$
$Z \perp\!\!\!\perp (\varepsilon_T, \varepsilon_M)$	$Z \perp\!\!\!\perp (\varepsilon_T, \varepsilon_Y)$	$Z \perp\!\!\!\perp (\varepsilon_T, \varepsilon_M, \varepsilon_Y)$

Notes: (*a*) *Model I* is the standard IV model enabling identification of the causal effect of T on M. *Model II* is the standard IV model enabling identification of the causal effect of T on Y. *Model III* is the IV Mediation Model with instrumental variable Z. (*b*) Panel A gives the DAG representation. Causal relationships are depicted by arrows, dashed lines denote statistical dependency, circles represent unobserved variables, and squares denote observed variables. Panel B presents the nonparametric structural equations. Conditioning variables are suppressed.

Under these conditions, the IV framework can be represented by three related models shown in Table 1. Model I treats the mediator as the outcome and identifies the causal effect of T on M. Model II treats the final outcome as the dependent variable and identifies the total effect of T on Y. Model III combines both relationships in a single mediation system.

Standard IV methods identify the treatment–mediator effect by treating M as the outcome (Model I), and the total effect of the treatment on the outcome (Model II). These correspond to the two margins along which the instrument generates variation. However, these results do not extend to the full mediation model (Model III). The causal effect of the mediator on the outcome—the parameter required to separate direct and indirect effects—is not identified. The reason is that all instrument-induced variation in the mediator operates through the treatment, so the data contain no independent variation in the mediator holding the treatment fixed. Without such variation, the direct effect of T on Y and the indirect effect through M cannot be disentangled. Thus, while the IV framework identifies the treatment–mediator effect and the total effect, it does not identify the decomposition of the total effect into direct and indirect components. In summary, the instrument shifts the treatment, and the treatment in turn shifts the mediator. As a result, all variation in the mediator induced by the instrument is mechanically tied to variation in the treatment.

Separating direct and indirect effects requires variation in the mediator that is not tied to the treatment. In the present setting, such variation is not observed. Any change in the mediator induced by the instrument necessarily reflects a change in the treatment, so it is impossible to isolate the effect of the mediator holding the treatment fixed. The next subsection formalizes this non-identification result in a linear framework.

2.4 Non-Identification in the Linear Model

We formalize the identification failure under linearity. The IV mediation model is:

$$Y = \alpha + \tau T + \theta M + X\beta + \varepsilon_Y, \tag{9}$$

$$M = \gamma_0 + \gamma T + X\gamma_2 + \varepsilon_M, \tag{10}$$

$$T = \pi_0 + Z\pi_1 + X\pi_2 + \varepsilon_T, \tag{11}$$

$$Z \perp\!\!\!\perp (\varepsilon_T, \varepsilon_M, \varepsilon_Y) \,\big|\, X, \tag{12}$$

where $\varepsilon_Y, \varepsilon_M, \varepsilon_T$ stand for unobserved error terms and the remaining Greek letters denote model coefficients. Two objects are identified under (9)–(12). First, since equations (10)–(11) constitute a standard IV model, the treatment-mediator effect γ is identified by 2SLS of M on T using Z as instrument:

$$\gamma = \frac{\mathrm{Cov}(Z, M \mid X)}{\mathrm{Cov}(Z, T \mid X)}.$$

Second, substituting (10) into (9) yields the reduced form

$$Y = \tilde{\alpha} + \tau^{\text{total}} T + X\tilde{\beta} + \tilde{\varepsilon}, \qquad \text{where} \quad \tau^{\text{total}} \equiv \tau + \theta\gamma \quad \text{and} \quad \tilde{\varepsilon} \equiv \varepsilon_Y + \theta\varepsilon_M. \tag{13}$$

Since $Z \perp\!\!\!\perp (\varepsilon_Y, \varepsilon_M) \mid X$ implies $Z \perp\!\!\!\perp \tilde{\varepsilon} \mid X$, the total effect τ^{total} is identified by 2SLS of Y on T:

$$\tau^{\text{total}} = \frac{\mathrm{Cov}(Z, Y \mid X)}{\mathrm{Cov}(Z, T \mid X)}.$$

Although γ and $\tau^{\text{total}} = \tau + \theta\gamma$ are identified, the direct effect τ and the mediator effect θ are not separately identified. To verify observational equivalence at the level of the IV moments, fix any alternative (τ', θ') with $\tau' + \theta'\gamma = \tau^{\text{total}}$ and define the implied outcome residual

$$\varepsilon'_Y \equiv \varepsilon_Y + (\tau - \tau')T + (\theta - \theta')M.$$

Substituting (10)–(11),

$$\varepsilon'_Y = \varepsilon_Y + (\theta - \theta')\,\varepsilon_M + \big[(\tau - \tau') + (\theta - \theta')\gamma\big]\big(\pi_0 + \pi_1 Z + X\pi_2 + \varepsilon_T\big) + (\theta - \theta')(\gamma_0 + X\gamma_2),$$

and the bracketed coefficient equals $(\tau + \theta\gamma) - (\tau' + \theta'\gamma) = 0$. Hence ε'_Y depends only on $(\varepsilon_M, \varepsilon_Y, X)$, so $\mathrm{Cov}(Z, \varepsilon'_Y \mid X) = 0$ and the parametrization (τ', θ', γ) satisfies the same IV moment conditions as (τ, θ, γ). Under $\gamma \neq 0$, the observational-equivalence family $\{(\tau', \theta') : \tau' + \theta'\gamma = \tau^{\text{total}}\}$ has both coordinates varying, so neither τ nor θ is separately identified.[1]

The source of the failure is that the instrument has no valid first stage for the mediator residual R: all variation in M that Z can explain operates through T, leaving no independent leverage to identify θ. No standard estimation strategy circumvents this problem. If T were exogenous, OLS of Y on T omitting M would recover τ^{total}, while including M would be inconsistent for (τ, θ) because $\mathrm{Cov}(\varepsilon_M, \varepsilon_Y \mid X) \neq 0$

[1] The degenerate case $\gamma = 0$ (no mediation channel) is excluded from the hypothesis because the indirect effect $\theta\gamma$ is identically zero, so $\tau = \tau^{\text{total}}$ is point-identified and only θ is unidentified (the IV moment $\mathrm{E}[Z(Y - \tau T - \theta M) \mid X] = 0$ loses its dependence on θ when $\mathrm{Cov}(Z, M \mid X) = \gamma\,\mathrm{Cov}(Z, T \mid X) = 0$). The proposition's claim of joint non-identification is the substantively interesting case in which a genuine mediation channel exists.

renders M endogenous—a bad control problem. When T is endogenous, 2SLS of Y on T using Z as instrument yields $\text{Cov}(Y, Z)/\text{Cov}(T, Z) = \tau^{\text{total}}$, since all covariation between Z and M operates through T. Estimating the structural equation (9) directly—2SLS of Y on (T, M) with Z as the sole instrument—is infeasible, as the first-stage predicted values $\hat{M}$ and $\hat{T}$ are collinear: all variation in $\hat{M}$ derives from $\hat{T}$. Identification of θ and τ therefore requires an additional assumption that breaks this collinearity in the instrument's moment conditions. The mediated confounding assumption of Section 3 does precisely this and provides identification.

3 The Mediated Confounding Assumption

We define the mediated confounding assumption (MCA) in Section 3.1 and compare it with sequential ignorability. Section 3.2 gives the graphical and structural interpretation. Section 3.3 derives the two operational consequences—conditional exogeneity and conditional relevance of the instrument for the mediator. Section 3.4 establishes equivalence to error-term independence in the linear model. Section 3.5 surveys alternative identification strategies and explains why the paper adopts MCA.

3.1 Definition and Formal Statement

All instrument-driven variation in M operates through T, leaving the instrument unable to disentangle the direct and indirect effects of T on Y (Section 2). We resolve this by imposing structure on the source of endogeneity.

Assumption 1 (Mediated Confounding Assumption)**.** *The unobserved determinants of treatment are independent of the unobserved determinants of the outcome:*

$$\{Y(t, m)\}_{(t,m)\in\mathcal{T}\times\mathcal{M}} \perp\!\!\!\perp \{T(z)\}_{z\in\mathcal{Z}} \mid X. \tag{14}$$

MCA (14) requires that the unobserved factors governing an individual's treatment response—how the individual's treatment varies with the instrument—be statistically independent of the unobserved factors governing the potential outcomes—what the outcome would be under any fixed combination of treatment and mediator. Any association between treatment-determining and outcome-determining unobservables must operate entirely through the mediator M.

Mediated confounding constrains the *source* of confounding, not its existence: both T and M remain endogenous. Specifically, $T \not\perp\!\!\!\perp (M(t), Y(t))$ and $M \not\perp\!\!\!\perp Y(t, m) \mid T$ continue to hold under MCA (14). This stands in sharp contrast with sequential ignorability (Imai et al., 2010), which assumes $Y(t', m), M(t) \perp\!\!\!\perp T \mid X$ and $Y(t', m) \perp\!\!\!\perp M \mid T, X$—effectively ruling out all unobserved confounding between T, M, and Y and rendering both T and M exogenous. Mediated confounding rules out only the unobserved confounding between T and Y that does not pass through M. In the language of potential outcomes, sequential ignorability restricts the mediator–outcome link; mediated confounding restricts the treatment–outcome link. This distinction carries practical consequences: the mediator is typically an early outcome subject to the same unobservables that drive the final outcome, making the mediator–outcome restriction difficult to justify. By contrast, mediated confounding leaves the mediator–outcome link unrestricted and imposes structure only on the relationship between treatment selection and outcome determination.

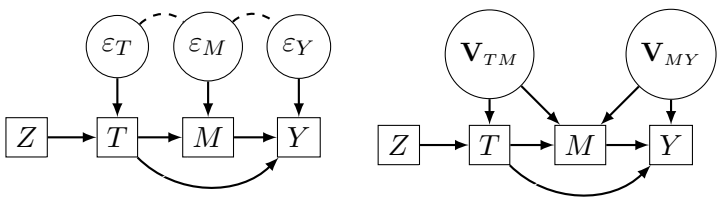
An Applied Illustration. Consider the effect of completing college (T) on health (Y) with health behaviors (M)—such as smoking, exercise, and diet—as mediator, using tuition discounts as instrument (Z). Mediated confounding requires that the factors governing an individual's college completion response to tuition discounts—such as family wealth, proximity to a college, and parental education—be independent of the factors governing her health potential at any fixed combination of education and behavior—such as genetic health endowment, local healthcare quality, and environmental exposures. This is plausible when the determinants of college completion are socioeconomic in nature while the determinants of baseline health are biological or environmental, and any socioeconomic–health link operates through the mediating behavior. The assumption would fail if, for example, individuals from low-wealth families both fail to complete college despite tuition discounts *and* have worse health through channels unrelated to the mediator—such as a direct effect of childhood poverty on chronic disease risk that is not captured by the health behavior measure M.

3.2 Graphical and Structural Interpretation

Table 2 presents two equivalent representations of the mediated confounding model. The left panel displays the DAG in terms of error terms: the key feature is the *absence* of the dashed line between ε_T and ε_Y—these error terms are independent—while the correlations $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_M$ and $\varepsilon_M \not\perp\!\!\!\perp \varepsilon_Y$ are preserved. Compared with Model III in Table 1, mediated confounding removes exactly one of the three pairwise error dependencies. The right panel offers a structural interpretation: in the unrestricted model, one can think of an unobserved confounding vector $\mathbf{V}$ simultaneously affecting T, M, and Y. MCA (14) splits $\mathbf{V}$ into two independent components $\mathbf{V} = [\mathbf{V}_{TM}, \mathbf{V}_{MY}]$, where $\mathbf{V}_{TM}$ drives (T, M) while $\mathbf{V}_{MY}$ drives (M, Y). This separation captures the essence of mediated confounding: unobserved confounders exist, but those driving treatment selection are distinct from those driving the outcome, with all confounding between treatment and outcome operating through the mediator.

Table 2: The Mediated Confounding IV Mediation Model

DAG Representation *Alternative Representation*



Notes: The DAG displays the mediated confounding IV mediation model. Compared to Model III in Table 1, the dashed line between ε_T and ε_Y is absent, indicating $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y$. Under this model, treatment T remains endogenous, mediator M remains endogenous conditional on T, but the instrument Z is exogenous with respect to $Y(t, m)$ conditional on T.

3.3 Key Properties: Conditional Exogeneity and Relevance

MCA (14) delivers two operational properties of the instrument. First, it generates a *new exogeneity condition*: the instrument is valid for the outcome conditional on the treatment. Second, it creates a *conditional relevance* channel: despite the exclusion restriction, the instrument becomes predictive of the mediator once we condition on the treatment:

Proposition 1 (New Instrument)**.** *Under the mediation IV conditions* (6)–(8) *and MCA* (14)*, the following conditions hold:*[2]

$$\textbf{\textit{Conditional Exogeneity:}} \quad Z \perp\!\!\!\perp Y(t,m) \,\big|\, (T,X) \qquad \textit{for all } (t,m) \in \mathcal{T} \times \mathcal{M}, \tag{15}$$

$$\textbf{\textit{Conditional Relevance:}} \quad Z \not\perp\!\!\!\perp M \,\big|\, (T,X). \tag{16}$$

The proof is in Appendix A.

Conditional exogeneity is the direct consequence of MCA (14). In the structural model of Table 2, conditioning on T as a collider on the path $Z \to T \leftarrow \varepsilon_T$ correlates the instrument with the treatment error ε_T. In the unrestricted model, this conditioning-induced dependence would propagate to the outcome through the error link $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_Y$, breaking the exogeneity of Z for Y given (T,X). MCA (14) closes precisely this channel by imposing $\varepsilon_T \perp\!\!\!\perp \varepsilon_Y \mid X$: the dependence between Z and ε_T that conditioning on T creates cannot reach ε_Y, so the instrument remains valid for the outcome conditional on the treatment, even though T itself remains endogenous.

Conditional relevance is a direct consequence of the collider effect and does not require MCA (14). In the structural decomposition of Table 2, the treatment T is a common effect of Z (through the first stage) and the confounding vector $\mathbf{V}_{TM}$. By Berkson's principle, conditioning on T—a collider on the path $Z \to T \leftarrow \mathbf{V}_{TM}$—induces a statistical dependence between Z and $\mathbf{V}_{TM}$. Since $\mathbf{V}_{TM}$ enters the mediator equation, this dependence transmits to M, making Z predictive of the residual variation in M given (T,X).

Remark 1 (The Inherited Instrument)**.** *The origin of the new instrument is best understood by considering the sub-DAG of Table 2 containing only M, Y, and the confounder $\mathbf{V}_{MY}$. In this reduced model, the unobserved confounder $\mathbf{V}_{TM}$ satisfies both conditions of a valid instrument for M: it shifts M through the $\mathbf{V}_{TM} \to M$ path (relevance) and is independent of $\mathbf{V}_{MY}$ under MCA* (14) *(exogeneity). Although $\mathbf{V}_{TM}$ is unobserved, the collider mechanism makes Z a proxy for $\mathbf{V}_{TM}$ conditional on T, allowing Z to inherit these instrumental properties.*

Proposition 1 establishes that under mediated confounding, the instrument Z satisfies both exogeneity and relevance for identifying the causal effect of M on Y, using T as a conditioning variable. Mediated confounding thus transforms a single treatment instrument into a valid instrument for both the treatment and the mediator.

3.4 Equivalence in the Linear Model

In the linear mediation model (9)–(12), the potential-outcome condition (14) reduces to a transparent restriction on the structural errors.

Proposition 2 (Linear Equivalence)**.** *In the linear mediation model* (9)–(12)*, MCA* (14) *is equivalent to conditional independence of the treatment and outcome errors:*

$$\varepsilon_T \perp\!\!\!\perp \varepsilon_Y \,\big|\, X. \tag{17}$$

[2]The proof in Appendix A invokes the graphoid axioms (symmetry, decomposition, weak union, contraction) of Dawid (1979), valid for any probability distribution admitting regular conditional distributions, as maintained by the standing regularity assumptions of Section 2.

The proof is in Appendix A.

Proposition 2 translates the potential-outcome condition into a single restriction on the structural errors: the unobservables governing treatment selection (ε_T) must be independent of those governing the outcome (ε_Y). In particular, $\text{Cov}(\varepsilon_T, \varepsilon_Y \mid X) = 0$. The remaining error correlations are unrestricted:

$$\text{Cov}(\varepsilon_M, \varepsilon_T \mid X) \neq 0 \quad \text{and} \quad \text{Cov}(\varepsilon_Y, \varepsilon_M \mid X) \neq 0. \tag{18}$$

Both conditions are required for the mediation problem to be nontrivial. The first generates the shared unobservable variation between T and M that powers the collider mechanism; the second ensures that the mediator remains endogenous with respect to the outcome. Neither is implied by MCA (14)—they are separate requirements on the error structure. A simple parametric example satisfying all three restrictions is $\varepsilon_M = \varepsilon + \alpha\varepsilon_T + (1-\alpha)\varepsilon_Y$, with $\alpha \in (0,1)$ and $\varepsilon, \varepsilon_T, \varepsilon_Y$ mutually independent.

To connect this representation to the identification argument above, we now show how the key properties of Section 3.3 arise in the linear model. Under (17) and (18), the conditional exogeneity (Proposition 1) and conditional relevance properties take an instructive algebraic form. For conditional exogeneity, the partial-covariance identity gives:

$$\text{Cov}(Z, \varepsilon_Y \mid X, T) = \underbrace{\text{Cov}(Z, \varepsilon_Y \mid X)}_{=0} - \text{Cov}(Z, T \mid X)\,\text{Var}(T \mid X)^{-1}\,\underbrace{\text{Cov}(T, \varepsilon_Y \mid X)}_{=\text{Cov}(\varepsilon_T, \varepsilon_Y \mid X) = 0} = 0, \tag{19}$$

so conditioning on T does not induce a spurious correlation between Z and ε_Y: the component of T driven by the instrument (via ε_T) is uncorrelated with the outcome error by (17).

For conditional relevance:

$$\text{Cov}(Z, \varepsilon_M \mid X, T) = \underbrace{\text{Cov}(Z, \varepsilon_M \mid X)}_{=0} - \text{Cov}(Z, T \mid X)\,\text{Var}(T \mid X)^{-1}\,\underbrace{\text{Cov}(T, \varepsilon_M \mid X)}_{=\text{Cov}(\varepsilon_T, \varepsilon_M \mid X) \neq 0} \neq 0, \tag{20}$$

since $\text{Cov}(Z, T \mid X) \neq 0$ by instrument relevance (3) and $\text{Cov}(\varepsilon_T, \varepsilon_M \mid X) \neq 0$ by (18). This is the collider mechanism in algebraic form: conditioning on T—a common effect of Z and ε_T—transmits the dependence $\varepsilon_T \not\perp\!\!\!\perp \varepsilon_M$ into a conditional dependence between Z and ε_M, making the instrument predictive of residual mediator variation.

3.5 Investigating Alternative Identification Strategies

The IV mediation model identifies the total effect of T on Y and the effect of T on M, but not the decomposition of the total effect into direct and indirect components using the instrument alone (Section 2). We therefore study which additional restrictions can identify the causal effect of M on Y while preserving the endogenous mediation problem.

We restrict attention to *single-instrument* identifying restrictions based on independence relations among the structural errors $(\varepsilon_T, \varepsilon_M, \varepsilon_Y)$. Any admissible restriction must be compatible with the maintained conditions (6)–(8), preserve treatment–mediator and mediator–outcome endogeneity, and identify the causal effect of M on Y using only the instrument Z.

Appendices B and C study a class of pairwise marginal and conditional independence restrictions. The analysis shows that most such restrictions fail to deliver identification while preserving the endogenous mediation structure. In particular, restrictions that eliminate treatment–mediator or mediator–outcome dependence remove the variation required for identification, while others fail to generate a valid

instrument for the mediator.

Within this class, exactly two restrictions satisfy the requirements above. 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 a control-function approach (see Appendix C).

We adopt MCA for three reasons. First, MCA does not require the additional assumptions commonly invoked by the control-function approach—inversion of the treatment equation through a strictly monotone reduced form, continuity of the endogenous treatment, or support conditions—and therefore remains available in settings with binary, discrete, or mixed treatments. Second, MCA is the more transparent restriction: it states that the unobserved determinants of treatment selection are independent of the unobserved determinants of the outcome, while leaving mediator–outcome endogeneity intact. Most importantly, MCA leads directly to the conditional IV setting described in 1. Thus, it benefits from the extensive econometric advances of the IV literature: monotonicity/separability conditions, 2SLS estimation, specification tests, and sensitivity analysis (Sections 5–6). The restriction $\varepsilon_M \perp\!\!\!\perp \varepsilon_Y \mid \varepsilon_T$ remains a valid alternative; Appendix C develops the control-function argument and discusses the additional structure it requires.

4 Identification and Estimation under Linearity

This section develops the identification argument and estimation theory under linearity. Under MCA (14) and the standard IV conditions, all four mediation parameters of the linear model (9)–(12)—the first-stage effect π_1, the treatment-mediator effect γ, the direct effect τ, and the mediator effect θ—are exactly identified by a four-equation GMM system (Theorem 1), where MCA yields the key moment condition $\mathrm{E}(\varepsilon_{i,T}\varepsilon_{i,Y}) = 0$ as an orthogonality restriction. The resulting total effect coincides with the standard Wald ratio (Corollary 1), and all causal parameters can be estimated by 2SLS regressions, confirming internal consistency. Section 6 extends these results to nonlinear settings.

4.1 Identification in the Linear Model

Throughout this section, all variables are orthogonalized with respect to $[\mathbf{1}, X]$—that is, we replace each variable by its residual from the OLS projection on the intercept and covariates. By the Frisch–Waugh–Lovell theorem, the slope coefficients in the linear mediation model (9)–(12) of Section 2.4 are unchanged, and the system reduces to:

$$T_i = \pi_1 \cdot Z_i + \varepsilon_{i,T}, \tag{21}$$

$$M_i = \gamma \cdot T_i + \varepsilon_{i,M}, \tag{22}$$

$$Y_i = \tau \cdot T_i + \theta \cdot M_i + \varepsilon_{i,Y}. \tag{23}$$

Observed data consists of n realizations $\{z_i, t_i, m_i, y_i\}_{i=1}^n$.

The IV conditions require that the instrument shifts treatment while remaining independent of the error terms:

$$Z_i \not\perp\!\!\!\perp T_i \quad \text{and} \quad Z_i \perp\!\!\!\perp (\varepsilon_{i,T}, \varepsilon_{i,M}, \varepsilon_{i,Y}). \tag{24}$$

MCA (14) imposes $\varepsilon_{i,T} \perp\!\!\!\perp \varepsilon_{i,Y}$ in the linear model (Proposition 2), and in particular $\mathrm{E}(\varepsilon_{i,T}\varepsilon_{i,Y}) = 0$, while permitting $\mathrm{E}(\varepsilon_{i,T}\varepsilon_{i,M}) \neq 0$ and $\mathrm{E}(\varepsilon_{i,M}\varepsilon_{i,Y}) \neq 0$. As established in Section 3.3, this generates conditional exogeneity ($Z_i \perp\!\!\!\perp \varepsilon_{i,Y} \mid T_i$) and conditional relevance ($Z_i \not\perp\!\!\!\perp M_i \mid T_i$).

The reduced-form outcome equation is obtained by substituting (22) into (23):

$$Y_i = \tau^{\text{total}} \cdot T_i + \eta_{i,Y}, \quad \text{where} \quad \eta_{i,Y} = \varepsilon_{i,Y} + \theta\varepsilon_{i,M} \quad \text{and} \quad \tau^{\text{total}} = \tau + \theta\gamma. \tag{25}$$

We now show that all structural parameters are identified under these conditions.

Theorem 1 (Identification of Mediation Parameters)**.** *Suppose conditions* (24)*, MCA* (17)*, and the rank conditions*

$$\sigma_{ZZ} > 0, \qquad \pi_1 \neq 0, \qquad \mathrm{cov}(\varepsilon_{i,T}, \varepsilon_{i,M}) \neq 0 \tag{26}$$

hold. Then the four parameters $\boldsymbol{\delta}_0 = [\pi_1, \gamma, \tau, \theta]$ are uniquely identified, with closed forms:

$$\pi_1 = \frac{\sigma_{ZT}}{\sigma_{ZZ}}, \tag{27}$$

$$\gamma = \frac{\sigma_{ZM}}{\sigma_{ZT}}, \tag{28}$$

$$\theta = \frac{\sigma_{ZT}\sigma_{TY} - \sigma_{TT}\sigma_{ZY}}{\sigma_{ZT}\sigma_{TM} - \sigma_{TT}\sigma_{ZM}}, \qquad \tau = -\frac{\sigma_{ZM}\sigma_{TY} - \sigma_{TM}\sigma_{ZY}}{\sigma_{ZT}\sigma_{TM} - \sigma_{TT}\sigma_{ZM}}, \tag{29}$$

where $\sigma_{AB} = \mathrm{E}(A_i B_i)$ for observed variables $A, B \in \{Z, T, M, Y\}$.

The three rank conditions in (26) have transparent content: $\sigma_{ZZ} > 0$ is instrument variation, $\pi_1 \neq 0$ is instrument relevance, and $\mathrm{cov}(\varepsilon_{i,T}, \varepsilon_{i,M}) \neq 0$ is treatment–mediator endogeneity—both required for the collider mechanism analyzed in Section 3. The proof is in Appendix D.

Corollary 1 (Total Effect Decomposition)**.** *The total effect decomposes as:*

$$\tau^{total} = \tau + \theta\gamma = \frac{\sigma_{ZY}}{\sigma_{ZT}}. \tag{30}$$

That is, the sum of the direct effect τ and the indirect effect $\theta\gamma$ equals the standard Wald ratio.

The proof is in Appendix D. Corollary 1 confirms internal consistency: the mediation decomposition recovers the same total effect as standard IV without invoking MCA.

Moment Conditions. Let $\boldsymbol{\delta}_0 = [\pi_1, \gamma, \tau, \theta]$ denote the vector of structural parameters and $\boldsymbol{w}_i = [Z_i, T_i, M_i, Y_i]$ the observed variables.[3] The model implies four orthogonality conditions:

$$\mathrm{E}[\boldsymbol{g}_i(\boldsymbol{w}_i|\boldsymbol{\delta}_0)] = \mathbf{0} \quad \text{where} \quad \boldsymbol{g}_i(\boldsymbol{w}_i|\boldsymbol{\delta}_0) \equiv \begin{pmatrix} Z_i \cdot (T_i - \pi_1 Z_i) \\ Z_i \cdot (M_i - \gamma T_i) \\ Z_i \cdot (Y_i - \tau T_i - \theta M_i) \\ (Y_i - \tau T_i - \theta M_i) \cdot (T_i - \pi_1 Z_i) \end{pmatrix} \tag{31}$$

The first three conditions exploit the exogeneity of Z with respect to each structural equation; they are standard IV moment conditions that hold in Model III of Table 1 regardless of MCA. The fourth

[3] $\boldsymbol{w}_i$ is assumed to be a stationary and ergodic stochastic process.

condition is the paper's key identifying restriction. It operationalizes MCA (17) as an orthogonality condition: the outcome residual $\varepsilon_{i,Y} = Y_i - \tau T_i - \theta M_i$ is orthogonal to the treatment residual $\varepsilon_{i,T} = T_i - \pi_1 Z_i$. This moment is what distinguishes the system from standard IV—the first three moments yield three equations in four unknowns; the fourth provides the additional restriction that pins down τ and θ separately. If MCA fails, the fourth moment has expectation $\text{Cov}(\varepsilon_{i,T}, \varepsilon_{i,Y}) \neq 0$, which biases $\hat{\theta}$ and $\hat{\tau}$ in a direction determined by the sign and magnitude of the direct treatment–outcome confounding.

Together, the four conditions yield an exactly identified system: four moment equations in four unknowns, with no overidentifying restrictions. Section 5 develops specification tests that become available when additional instruments are present.

4.2 Two-Stage Least Squares Estimation

Each identification equation in Theorem 1 corresponds to a standard econometric regression, and the sample-analogue GMM estimator replaces population covariances by their sample counterparts.[4] The first-stage effect π_1 is estimated by OLS of T on Z. The treatment-mediator effect γ is estimated by 2SLS of M on T, using Z as instrument.

The mediator-outcome and direct effects (θ, τ) are estimated by a conditional 2SLS regression of Y on M and T, using Z as instrument for M conditional on T:

$$\text{First Stage:} \quad M_i = \delta_1 Z_i + \delta_2 T_i + \eta_{i,M}, \tag{32}$$

$$\text{Second Stage:} \quad Y_i = \theta \cdot \hat{M}_i + \tau \cdot T_i + \varepsilon_{i,Y}. \tag{33}$$

Including T in the first stage is the estimation analogue of the collider mechanism: the regression isolates the component of Z's predictive power for M that operates through the induced Z–$\varepsilon_{i,M}$ dependence conditional on T, rather than through the causal chain $Z \to T \to M$.

The conditional 2SLS estimator for θ admits a compact representation. Since all variables have been orthogonalized with respect to $[\mathbf{1}, X]$, the Frisch–Waugh–Lovell projection onto T alone suffices. Define $\tilde{Y} \equiv M_T Y$, $\tilde{M} \equiv M_T M$, and $\tilde{Z} \equiv M_T Z$, where $M_T \equiv I_n - T(T'T)^{-1}T'$ is the orthogonal projection onto the complement of the column space of T. Then:

$$\hat{\theta} = \frac{\tilde{Z}'\tilde{Y}}{\tilde{Z}'\tilde{M}}. \tag{34}$$

Conditional relevance ensures $\tilde{Z}'\tilde{M} \neq 0$ asymptotically, and conditional exogeneity (Proposition 1) ensures $\text{E}(\tilde{Z}'\varepsilon_{i,Y}) = 0$, so $\hat{\theta}$ is consistent for θ. The strength of the conditional first stage—the F-statistic from the regression of M on Z and T—is a key diagnostic: because the conditional relevance channel operates through the collider mechanism, it may be substantially weaker than the unconditional first stage for T, even when the latter is strong. The following remark formalizes this relationship.

Remark 2 (Conditional First-Stage Strength)**.** *Let $\rho_{TM} \equiv Corr(\varepsilon_{i,T}, \varepsilon_{i,M})$ and $r_{TZ} \equiv Corr(T_i, Z_i)$. In the linear model (21)–(22), the partial correlation of M and Z conditional on T depends only on the*

[4] A reference Python implementation, `iv_mediation_applied.py`, is available with the replication materials. The package implements the conditional 2SLS estimator of this section, the specification tests and κ-based sensitivity analysis of Section 5, the Olea–Pflueger (Montiel Olea and Pflueger, 2013) effective F-statistic of Remark 2, the quadruple-LASSO post-double selection of Appendix H, and supports analytic and probability weights, one- and two-way fixed effects, classical, heteroskedasticity-robust, cluster-robust, and Conley spatial standard errors, and pairs and block bootstrap inference.

error correlation ρ_{TM} and the instrument-treatment correlation r_{TZ}:

$$r_{MZ|T} = \frac{-\rho_{TM}\, r_{TZ}}{\sqrt{1 - \rho_{TM}^2\,(1 - r_{TZ}^2)}}. \tag{35}$$

The standard first-stage F-statistic of T on Z, the conditional F-statistic for M on (Z, T) and their ratio take the closed form:

$$F_{TZ} = (n-2)\,\frac{r_{TZ}^2}{1 - r_{TZ}^2},\quad F_{MZ|T} = (n-3)\,\frac{r_{TZ}^2\,\rho_{TM}^2}{1 - \rho_{TM}^2},\quad \textit{and}\quad \frac{F_{MZ|T}}{F_{TZ}} \xrightarrow{n\to\infty} \frac{\rho_{TM}^2\,(1 - r_{TZ}^2)}{1 - \rho_{TM}^2}.$$

Two implications follow. First, the conditional first stage vanishes when $\rho_{TM} = 0$: without treatment-mediator confounding, the collider mechanism has no power, and the mediator effect θ is not identified—confirming the rank condition in Theorem 1. Second, even when F_{TZ} is large, $F_{MZ|T}$ may be small: the F-ratio is bounded by $\rho_{TM}^2/(1-\rho_{TM}^2)$ and decreases in r_{TZ}^2. For example, with $|\rho_{TM}| = 0.3$ and $r_{TZ} = 0.5$, the asymptotic ratio is approximately 0.068, so a standard first-stage F_{TZ} of 100 corresponds to a conditional first-stage $F_{MZ|T}$ of only 6.8. Researchers should report the conditional first-stage strength alongside the standard first stage. Because the conditional regression $M_i = \delta\, Z_i + \xi\, T_i + \nu_i$ generally exhibits heteroskedasticity—the residual variance depends on T_i through the partialling-out step, and on cluster structure in panel applications—the appropriate diagnostic is the effective F-statistic of Montiel Olea and Pflueger (2013)—denoted F_{eff}, the first-stage F computed with a heteroskedasticity-robust (or cluster-robust) variance estimator for $Z'\nu$. The Staiger–Stock benchmark of $F > 10$ is a conservative i.i.d. special case; F_{eff} should be compared to the critical values tabulated in Montiel Olea and Pflueger (2013) for the chosen worst-case bias tolerance τ_{bias} (typically $\tau_{\text{bias}} = 10\%$ of the OLS benchmark). When F_{eff} falls below the relevant critical value, weak-instrument-robust inference—Anderson–Rubin or conditional likelihood-ratio confidence intervals—should be reported in place of conventional 2SLS standard errors.

Remark 3 (Asymptotic Distribution)**.** *Under standard regularity conditions (i.i.d. sampling, finite fourth moments, correct specification of (31)), the GMM estimator $\hat{\boldsymbol{\delta}} = [\hat{\pi}_1, \hat{\gamma}, \hat{\tau}, \hat{\theta}]$ is consistent and asymptotically normal with covariance matrix $(\boldsymbol{G}'\boldsymbol{\Sigma}_{\boldsymbol{g}}^{-1}\boldsymbol{G})^{-1}$, consistently estimated by the sandwich formula; see Appendix D. Standard errors for the indirect effect $\theta\gamma$ follow from the delta method or bootstrap. The i.i.d. assumption can be relaxed to accommodate conditional heteroskedasticity (Hayashi, 2000).*

Remark 4 (Treatment-Dependent Indirect Effects)**.** *The linear model (23) restricts the natural indirect effect (5) to be constant across treatment levels: NIE$(t) = \theta\gamma$ for all t. Appendix E relaxes this by augmenting the outcome equation with an interaction term $\rho \cdot T_i \cdot M_i$, yielding NIE$(t) = (\theta + \rho t)\gamma$. The interaction parameter ρ is exactly identified under MCA since the product Z_iT_i serves as an additional excluded instrument because MCA implies $\mathrm{E}(Z_iT_i \cdot \varepsilon_{i,Y}) = 0$. The model can be estimated by 2SLS, regressing Y on (M, MT), using (Z, ZT) as instruments for (M, MT), with $(1, T)$ as exogenous regressors. The rank condition for identifying (τ, θ, ρ)—necessary and sufficient—is that the 3×3 cross-moment matrix between the instruments (Z, T, ZT) and the endogenous regressors (T, M, TM) be nonsingular (Appendix E, eq. (114)).*

5 Sensitivity Analysis and Test Limitations under Linearity

MCA (14) is the restriction that separates the direct effect of the treatment from the mediated effect operating through M. It is therefore natural to ask whether the assumption can be tested. The central result of this section is negative: with treatment instruments alone, MCA is not testable. The MCA restriction sets the latent covariance $\mathrm{E}[\varepsilon_T \varepsilon_Y]$ to zero, whereas the overidentifying moments generated by treatment instruments are Z-orthogonality restrictions. These moments can diagnose failures of the maintained IV model—exogeneity, exclusion, linearity, or parameter stability across values of Z—but they do not, by themselves, reveal the latent dependence ruled out by MCA.

The source of the difficulty is simple. Under mediation exclusion, any instrument-induced movement in the mediator is inherited from the treatment first stage:

$$\mathrm{cov}(\boldsymbol{Z}_i, M_i) = \gamma \, \mathrm{cov}(\boldsymbol{Z}_i, T_i).$$

Thus the reduced-form effects of $\boldsymbol{Z}_i$ on T_i and M_i span a single direction. A violation of MCA can then be absorbed into the probability limits of the direct and mediator-effect coefficients without violating the Z-based moment conditions used by standard overidentification tests.

The constructive implication is that the feasible diagnostic in the canonical design is sensitivity analysis. We index departures from MCA by

$$\kappa \equiv \mathrm{cov}(\varepsilon_T, \varepsilon_Y),$$

the scalar covariance that MCA sets to zero. Subsection 5.1 first distinguishes apparent overidentification, which is generated by transformations or discretizations of treatment instruments, from informative overidentification, which requires independent mediator variation. It then explains why the LR construction remains useful in the main text: in designs with a genuine mediator instrument, unrestricted GMM estimates the same κ that the sensitivity analysis varies in the canonical design. Subsection 5.2 develops the resulting κ-based sensitivity framework.

5.1 Apparent versus Informative Overidentification

Testing MCA requires more than a surplus of moments. It requires an instrument set that shifts the mediator in a direction not spanned by the treatment first stage. Denoting the instrument vector by $\boldsymbol{Z}_i$, the relevant rank condition is

$$\mathrm{rank}\Big(\mathrm{cov}(\boldsymbol{Z}_i, T_i) \quad \mathrm{cov}(\boldsymbol{Z}_i, M_i)\Big) = 2. \tag{36}$$

Equivalently, there must exist a linear combination $a'\boldsymbol{Z}_i$ such that $\mathrm{cov}(a'\boldsymbol{Z}_i, T_i) = 0$ and $\mathrm{cov}(a'\boldsymbol{Z}_i, M_i) \neq 0$. This is the precise sense in which overidentification must be informative for MCA: after removing the treatment channel, some instrument variation must remain in the mediator equation.

This condition fails when every instrument affects the mediator only through the treatment. In that case $M_i = \gamma T_i + \varepsilon_{i,M}$ with $\mathrm{E}[\boldsymbol{Z}_i \varepsilon_{i,M}] = \mathbf{0}$, so

$$\mathrm{cov}(\boldsymbol{Z}_i, M_i) = \gamma \, \mathrm{cov}(\boldsymbol{Z}_i, T_i),$$

and the matrix in (36) has rank one. Nonlinear transformations of a scalar instrument, such as $(Z_i, Z_i^2, Z_i^3, \ldots)$, or indicators for a multi-valued instrument, can create additional moments.[5] For a K-dimensional instrument vector, the moment conditions in (31) yield $3K+1$ moments for $K+3$ parameters, and hence $2(K-1)$ overidentifying restrictions. These are, however, only apparent overidentifying restrictions for MCA when the rank condition fails. They may be useful omnibus diagnostics for the maintained linear IV specification, but they do not test the latent covariance restriction.

Observational equivalence. The non-testability result can be seen directly. Suppose every component of $\boldsymbol{Z}_i$ affects the mediator only through treatment, so $M_i = \gamma T_i + \varepsilon_{i,M}$ with $\mathrm{E}[\boldsymbol{Z}_i \varepsilon_{i,M}] = \mathbf{0}$. Let

$$\kappa \equiv \mathrm{cov}(\varepsilon_T, \varepsilon_Y), \qquad \rho_{TM} \equiv \mathrm{cov}(\varepsilon_T, \varepsilon_M) \neq 0.$$

For any value of κ, define the shifted coefficients

$$\theta^* \equiv \theta + \frac{\kappa}{\rho_{TM}}, \qquad \tau^* \equiv \tau - \gamma \frac{\kappa}{\rho_{TM}}. \tag{37}$$

Substituting the structural equations gives

$$Y_i - \tau^* T_i - \theta^* M_i = \varepsilon_{i,Y} - \frac{\kappa}{\rho_{TM}} \varepsilon_{i,M}.$$

Therefore the Z-orthogonality moments and the MCA moment both hold at (τ^*, θ^*):

$$\mathrm{E}[\boldsymbol{Z}_i \cdot (Y_i - \tau^* T_i - \theta^* M_i)] = \mathbf{0}, \qquad \mathrm{E}[T_i \cdot (Y_i - \tau^* T_i - \theta^* M_i)] = 0. \tag{38}$$

The first equality follows from IV exogeneity, $\mathrm{E}[\boldsymbol{Z}_i \varepsilon_{i,Y}] = \mathrm{E}[\boldsymbol{Z}_i \varepsilon_{i,M}] = \mathbf{0}$. The second reduces to

$$\mathrm{cov}(\varepsilon_T, \varepsilon_Y) - \frac{\kappa}{\rho_{TM}} \mathrm{cov}(\varepsilon_T, \varepsilon_M) = \kappa - \kappa = 0.$$

Thus a model with $\kappa \neq 0$ is observationally equivalent, in the restricted moments, to a model satisfying MCA but with coefficients (τ^*, θ^*). The 2SLS probability limits are $\mathrm{plim}(\hat{\theta}) = \theta^*$ and $\mathrm{plim}(\hat{\tau}) = \tau^*$, and the violation of MCA is absorbed into the estimated effects rather than appearing as a population moment failure.

J- and t-tests. The preceding argument determines the interpretation of the standard GMM diagnostics. The Hansen J-statistic (139) tests the full moment vector in (31) jointly and rejects when at least one moment fails. The targeted t-statistic (140) evaluates the MCA moment $\mathrm{E}[T_i \varepsilon_{i,Y}] = 0$ using the GMM-implied standard error of the evaluated moment. Their formulas and limiting distributions are stated in Appendix F. When the rank condition (36) fails, neither statistic has asymptotic power against MCA violations of the form described above: the restricted moments hold in population at the shifted coefficients. The tests should therefore be read as omnibus diagnostics for the maintained linear IV model, not as tests of MCA.

[5] Using nonlinear functions such as Z^2 as overidentifying instruments follows Dieterle and Snell (2016), who establish validity under mean independence $\mathrm{E}[\varepsilon \mid Z] = 0$, a condition implied by the stronger independence $Z \perp\!\!\!\perp (\varepsilon_T, \varepsilon_M, \varepsilon_Y)$ maintained in the paper. The construction also presumes that the linear specification (21)–(23) is correct; if the true model is nonlinear, the additional moments may fail for reasons unrelated to MCA.

The LR bridge. The LR construction differs from the J- and t-tests because its unrestricted model introduces the same scalar κ that indexes the sensitivity analysis. The restricted model imposes $\kappa = 0$, whereas the unrestricted system replaces the MCA moment by

$$(Y_i - \tau T_i - \theta M_i) \cdot (T_i - \boldsymbol{\pi}_1' \boldsymbol{Z}_i) - \kappa. \tag{39}$$

When the rank condition (36) holds, κ is identified by the unrestricted GMM system. The LR statistic compares the restricted and unrestricted GMM objective functions, using the unrestricted weighting matrix:

$$\text{LR} = n\Big(\bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\delta}})' \hat{\boldsymbol{\Sigma}}_{\boldsymbol{g},n}^{u\,-1} \bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\delta}}) \;-\; \bar{\boldsymbol{g}}_n^u(\hat{\boldsymbol{\delta}}_n^u)' \hat{\boldsymbol{\Sigma}}_{\boldsymbol{g},n}^{u\,-1} \bar{\boldsymbol{g}}_n^u(\hat{\boldsymbol{\delta}}_n^u)\Big) \xrightarrow{d} \chi_1^2 \tag{40}$$

under MCA. The unrestricted estimate $\hat{\kappa}$ and its GMM sandwich standard error then provide a point estimate and confidence interval for the departure from mediated confounding.[6]

This is the reason to retain the LR construction in the main text. In an informative overidentified design, $\hat{\kappa}$ can be used as the input to the bias-corrected mediator-effect estimator

$$\hat{\theta}_{\text{adj}}(\kappa_0) = \hat{\theta} - \frac{\kappa_0}{\widehat{\text{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)}. \tag{41}$$

When $\kappa_0 = \hat{\kappa}$, this adjustment connects the unrestricted overidentified model to the sensitivity analysis developed below. In the canonical treatment-instrument design, by contrast, $\hat{\kappa}$ is not separately identified: the unrestricted LR system cannot distinguish κ from the coefficient shift in (37). The same adjustment is then used with hypothetical or breakeven values of κ_0, rather than with a plug-in estimate.

Mediator-instrument designs. This limitation is specific to treatment instruments. In two-instrument mediation designs, such as Frölich and Huber (2017b), a second instrument may enter the mediator equation directly,

$$M_i = \gamma T_i + \phi Z_{2,i} + \varepsilon_{i,M}, \qquad \phi \neq 0,$$

while remaining excluded from the outcome equation. Then $\text{cov}(\boldsymbol{Z}_i, M_i)$ need not be proportional to $\text{cov}(\boldsymbol{Z}_i, T_i)$, the rank condition (36) can hold, and unrestricted GMM can identify the violation parameter κ. The contrast is useful because it shows that the non-testability result is not a generic limitation of GMM tests, but a consequence of relying only on treatment instruments. In the canonical design studied here, κ is not identified, so Section 5.2 uses breakeven sensitivity analysis instead.

5.2 Sensitivity Analysis for the Mediator Effect

Because MCA is not testable with treatment instruments alone, the relevant empirical question is how large a departure from MCA would be required to overturn the substantive conclusion. The sensitivity analysis targets the identifying restriction directly by working with $\kappa = \text{cov}(\varepsilon_T, \varepsilon_Y)$. In the canonical just-identified case, κ is not identified and must be calibrated. The advantage of this formulation is that the violation is one-dimensional: the researcher assesses a single structural covariance, rather than a pair

[6] The LR test is analogous to the Durbin–Wu–Hausman test for endogeneity (Hausman, 1978), but it targets the specific covariance $\text{cov}(\varepsilon_T, \varepsilon_Y)$ that MCA sets to zero, rather than overall regressor exogeneity.

of reduced-form partial R^2 parameters.[7]

5.2.1 Direct Sensitivity Analysis via κ

The following proposition gives the bias induced by an MCA violation and the corresponding breakeven and bias-adjusted estimators.

Proposition 3 (κ-Based Sensitivity)**.** *Maintain the linear model* (21)–(23)*, instrument exogeneity* $\operatorname{cov}(Z, \varepsilon_T) = \operatorname{cov}(Z, \varepsilon_M) = \operatorname{cov}(Z, \varepsilon_Y) = 0$*, and the rank conditions* $\sigma_{ZZ} > 0$*,* $\pi_1 \neq 0$*, and* $\operatorname{cov}(\varepsilon_T, \varepsilon_M) \neq 0$*. Allow* $\kappa \equiv \operatorname{cov}(\varepsilon_T, \varepsilon_Y) \neq 0$*. Then:*

(a) *The asymptotic bias of the 2SLS estimator* $\hat{\theta}$ *from* (34) *is*

$$\operatorname{plim}(\hat{\theta}) - \theta = \frac{\kappa}{\operatorname{cov}(\varepsilon_T, \varepsilon_M)}. \tag{42}$$

(b) *Let* $\hat{\varepsilon}_T = T - \hat{\pi}_1 Z$ *and* $\hat{\varepsilon}_M = M - \hat{\gamma}_{\text{IV}} T$*, where* $\hat{\gamma}_{\text{IV}} = \widehat{\operatorname{cov}}(Z, M)/\widehat{\operatorname{cov}}(Z, T)$*. Then* $\widehat{\operatorname{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$ *is consistent for* $\operatorname{cov}(\varepsilon_T, \varepsilon_M)$ *for any value of* κ*.*

(c) *Under MCA* ($\kappa = 0$)*, the breakeven value* $\hat{\kappa}^* = \hat{\theta} \cdot \widehat{\operatorname{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$ *is consistent for* $\kappa^* \equiv \theta \operatorname{cov}(\varepsilon_T, \varepsilon_M)$*, and* $\sqrt{n}(\hat{\kappa}^* - \kappa^*)$ *is asymptotically normal by the delta method. More generally,* $\hat{\kappa}^* \xrightarrow{p} \kappa^* + \kappa$*.*

(d) *When* $\kappa_0 = \kappa$*, the adjusted estimator* $\hat{\theta}_{\text{adj}}(\kappa_0)$ *is* $\sqrt{n}$*-consistent for* θ*, and the confidence interval* $\text{CI}_{\kappa_0}(\theta)$ *defined in* (46) *has asymptotic coverage* $1 - \alpha$ *under MCA. For* $\kappa \neq 0$*, the interval is approximately valid to first order when estimation uncertainty in* $\widehat{\operatorname{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$ *is small relative to the sampling variability of* $\hat{\theta}$*; exact coverage requires replacing* $\widehat{\text{SE}}(\hat{\theta})$ *with the delta-method standard error of the full adjusted estimator.*

The proof is in Appendix F.

Breakeven values and adjusted inference. The signed breakeven covariance is

$$\hat{\kappa}^* \equiv \hat{\theta} \cdot \widehat{\operatorname{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M). \tag{43}$$

It is the value of $\operatorname{cov}(\varepsilon_T, \varepsilon_Y)$ that sets the adjusted mediator effect to zero. Hence $|\hat{\kappa}^*|$ is the minimum magnitude of treatment–outcome error covariance needed to overturn the point estimate, and $\operatorname{sign}(\hat{\kappa}^*)$ gives the required direction of confounding. The analogous inference breakeven is

$$\hat{\kappa}^*_{\text{CI}} \equiv \left(\hat{\theta} - z_{\alpha/2}\widehat{\text{SE}}(\hat{\theta})\right) \cdot \widehat{\operatorname{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M), \tag{44}$$

the smallest value of κ, in magnitude and in the relevant direction, that brings the $(1 - \alpha)$ confidence interval for θ to include zero. For a hypothesized violation κ_0, the adjusted estimate and confidence

[7] A complementary R^2-based framework adapted from Cinelli and Hazlett (2024) assesses robustness of the conditional reduced form to omitted confounders affecting both the residualized instrument $\tilde{Z}$ and the residualized outcome $\tilde{Y}$. Because this framework targets a distinct threat—confounders operating outside the treatment-equation channel—it does not test MCA. We present it in Appendix F and recommend reporting both diagnostics when feasible.

interval are

$$\hat{\theta}_{\text{adj}}(\kappa_0) = \hat{\theta} - \frac{\kappa_0}{\widehat{\text{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)}, \tag{45}$$

$$\text{CI}_{\kappa_0}(\theta) = \hat{\theta}_{\text{adj}}(\kappa_0) \pm z_{\alpha/2}\widehat{\text{SE}}(\hat{\theta}). \tag{46}$$

By Proposition 3, the adjusted estimator is consistent when $\kappa_0 = \kappa$, and the interval has exact nominal coverage under MCA. Away from MCA, $\widehat{\text{SE}}(\hat{\theta})$ omits estimation uncertainty in $\widehat{\text{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)$; the exact first-order variance is obtained by applying the delta method to the full adjusted estimator. In the simulations below, this simpler interval has good finite-sample performance across the violation magnitudes considered.

Connection to conditional first-stage strength. The denominator in (45) is the same covariance that governs the conditional first-stage strength (Remark 2). When $\text{cov}(\varepsilon_T, \varepsilon_M)$ is close to zero, a small value of κ induces a large bias in $\hat{\theta}$, and the breakeven $|\hat{\kappa}^*|$ is correspondingly small. Thus a weak conditional first stage signals two concerns simultaneously: imprecision in the mediator-effect estimate and high sensitivity to departures from MCA. Conversely, a strong conditional first stage implies that only a substantively large treatment–outcome error covariance can overturn the estimate.

Interpretation and benchmarking. The quantity $\hat{\kappa}^*$ has a direct structural interpretation, but its scale may be unfamiliar. Three benchmarks are useful in applications. First, when an informative overidentified design is available, $\hat{\kappa}^*$ can be compared with the LR estimate $\hat{\kappa}$. Second, it can be compared with observed covariances $\widehat{\text{cov}}(\hat{\varepsilon}_T, X_j)$ between the treatment-equation residual and observed outcome determinants X_j; this asks whether an omitted determinant would need to be more strongly related to the treatment residual than observed covariates are. Third, it can be expressed as the structural error correlation in Remark 6, which is bounded by one in absolute value. In the canonical just-identified design, these benchmarks replace a formal test of MCA.

Remark 5 (Nesting of Tests and Sensitivity Analysis)**.** *When the rank condition* (36) *holds, the unrestricted LR/GMM system of Section 5.1 consistently estimates κ. Substituting $\kappa_0 = \hat{\kappa}$ into* (45) *gives*

$$\hat{\theta}_{\text{adj}}(\hat{\kappa}) = \hat{\theta} - \frac{\hat{\kappa}}{\widehat{\text{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)}. \tag{47}$$

This estimator is consistent for θ whether or not MCA holds. Indeed, Proposition 3(a)–(b) gives $\text{plim}(\hat{\theta}) = \theta + \kappa/\rho_{TM}$ and $\text{plim}\{\widehat{\text{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)\} = \rho_{TM}$, while unrestricted GMM gives $\text{plim}(\hat{\kappa}) = \kappa$. The two bias terms therefore cancel. The probability limit coincides with that of the unrestricted GMM estimator $\hat{\theta}^u$ that treats κ as a free parameter in the full system (31)–(39).[8]

Thus the LR analysis and the sensitivity analysis are nested. In informative overidentified designs, κ is estimated and can be used for bias correction. In the canonical treatment-instrument design, κ is not identified, and the researcher varies κ_0 over substantively plausible or breakeven values.

[8] Both estimators are $\sqrt{n}$-consistent. Under efficient two-step GMM, they need not have identical asymptotic variance because $\hat{\theta}^u$ exploits the full joint moment structure whereas $\hat{\theta}_{\text{adj}}(\hat{\kappa})$ is a plug-in correction. Inference after estimating κ should therefore use the unrestricted GMM sandwich variance.

Remark 6 (Breakeven as a Structural Correlation)**.** *The breakeven can be reported in correlation units:*

$$\hat{\rho}^*_{TY} \equiv \frac{\hat{\kappa}^*}{\hat{\sigma}_{\varepsilon_T}\hat{\sigma}_{\varepsilon_Y}} = \frac{\hat{\theta} \cdot \widehat{\mathrm{cov}}(\hat{\varepsilon}_T, \hat{\varepsilon}_M)}{\hat{\sigma}_{\varepsilon_T}\hat{\sigma}_{\varepsilon_Y}}, \tag{48}$$

where $\hat{\sigma}_{\varepsilon_T}$ and $\hat{\sigma}_{\varepsilon_Y}$ are computed from the first-stage and outcome-equation residuals under MCA.[9] *Although $\hat{\rho}^*_{TY}$ is not itself a sample correlation, the bound $|\mathrm{Corr}(\varepsilon_T, \varepsilon_Y)| \leq 1$ gives an immediate benchmark. If $|\hat{\rho}^*_{TY}| > 1$, no feasible error correlation can reduce the estimated mediator effect to zero. Values near one indicate that only very strong latent treatment–outcome dependence would overturn the estimate; values near zero indicate fragility.*

5.2.2 Sensitivity of the Mediation Decomposition

The indirect effect $\hat{\theta}\hat{\gamma}$ inherits sensitivity from both of its components, but the relevant threats differ across components. The treatment–mediator effect $\hat{\gamma}$ is identified by standard IV using Z as an instrument for T, so its sensitivity is assessed using the Cinelli and Hazlett (2024) framework applied to the M-on-T IV regression, targeting violations of instrument exogeneity such as $\mathrm{cov}(Z, \varepsilon_M) \neq 0$. The mediator–outcome effect $\hat{\theta}$ is sensitive to MCA violations, addressed by the κ-based analysis above, and to conditional reduced-form confounding, addressed by the R^2-based framework in Appendix F. Since the indirect effect is overturned whenever either component is overturned, its robustness is bounded above by the weaker of the component robustness values.

6 Extensions to Nonlinear Settings

This section extends the mediated confounding framework beyond linearity and makes four points. First, we extend the Local Instrumental Variable (LIV) model of Heckman and Vytlacil (1999, 2005) to a mediation setting. Second, under the resulting LIV representation, we identify the *Marginal Mediator Effect* (MME), the effect of switching the mediator at a given value of latent mediator resistance. Unlike the standard Marginal Treatment Effect (MTE), mediation requires a two-step margin: a first derivative isolates the treatment-resistance margin, and a second derivative maps that margin into the mediator-resistance margin. Third, when the support of Z collapses to two points, the derivative formulas have a finite-difference analogue: conditional Wald ratios recover local mediator effects under an additional monotonicity condition tailored to the binary-support case. Fourth, the framework delivers testable implications in the form of falsification tests that probe the mediation-specific identifying assumptions.

6.1 The General Nonparametric Framework

This subsection states the nonlinear mediation-IV environment under MCA. Let the observed variables be (Y, T, M, Z, X). Denote by $Y(t, m)$ the potential outcome under $(T, M) = (t, m)$, write $Y(t) \equiv Y(t, M(t))$, let $M(t)$ denote the potential mediator under treatment state t, and let $T(z)$ denote the potential treatment under instrument value z. The key objects for the rest of the section are the latent treatment resistance U_T, the Marginal Mediator Response (MMR) $\rho(t, u)$, and the MME, which

[9] Under MCA, $\hat{\theta}$ is consistent for θ, so the residual $Y - \hat{\tau}T - \hat{\theta}M - \hat{\beta}_X X$ is consistent for ε_Y. The value $\hat{\rho}^*_{TY}$ is therefore a threshold computed under the maintained null: it is the hypothetical error correlation required to overturn the estimate.

measures the causal effect of moving the mediator from 0 to 1 at mediator resistance u while holding treatment fixed at t. These objects are formally defined in the subsections that follow.

The mediation IV conditions (6)–(8) and the mediated confounding assumption (MCA (14)) remain in force. As established in Section 3.3, these conditions jointly imply conditional exogeneity—$Z \perp\!\!\!\perp Y(t,m) \mid (T,X)$—and conditional relevance—$Z \not\perp\!\!\!\perp M \mid (T,X)$—which render Z a valid instrument for M in the outcome equation upon conditioning on T. The conditional relevance exploits the endogeneity of T with respect to M: conditioning on T creates a collider, through which Z inherits dependence with the unobserved factors driving both T and M (Remark 1). The resulting instrument-driven variation in type composition within treatment strata is termed *compositional variation*.

6.2 Identification via the Local Instrumental Variable Approach

We begin with the case in which Z is continuous. Consider binary T and M, each taking values in $\{0,1\}$. The structural model follows Table 2: $T = f_T(Z,\nu_T)$, $M = f_M(T,\nu_T,\nu_Y)$, and $Y = f_Y(T,M,\nu_Y,\epsilon)$. Here ν_T and ν_Y are unobserved random variables and ϵ is an idiosyncratic outcome shock.[10] The IV exogeneity condition is $Z \perp\!\!\!\perp (\nu_T,\nu_Y,\epsilon)$, and the mediated confounding assumption (MCA (14)) takes the form $\nu_T \perp\!\!\!\perp (\nu_Y,\epsilon)$ in this structural model. These jointly imply a key property:

$$(Z,\nu_T,T) \perp\!\!\!\perp (Y(t,m),\nu_Y). \tag{49}$$

Property (49) holds because Z,ν_T are jointly independent of (ν_Y,ϵ), $T = f_T(Z,\nu_T)$ is a function of (Z,ν_T), and $Y(t,m) = f_Y(t,m,\nu_Y,\epsilon)$ is a function of (ν_Y,ϵ).

Treatment Equation. Following Heckman and Vytlacil (2005), we assume that the treatment equation is separable:

$$T = \mathbf{1}[\zeta(Z) \geq \phi(\nu_T)], \tag{50}$$

where $\phi(\nu_T)$ is absolutely continuous. Applying the probability integral transform to both sides of the inequality yields the propensity score representation:

$$T = \mathbf{1}[P_T(Z) \geq U_T], \tag{51}$$

where $P_T(Z) = \Pr(T=1 \mid Z)$ is the propensity score and $U_T = F_{\phi(\nu_T)}(\phi(\nu_T)) \sim \text{Uniform}[0,1]$. We refer to U_T as the *treatment resistance*: higher values correspond to lower selection into treatment for any given propensity. The exogeneity $Z \perp\!\!\!\perp \nu_T$ implies $P_T(Z) \perp\!\!\!\perp U_T$.

Identification of Counterfactual Mediator. The first step recovers the counterfactual mediator $M(t) = f_M(t,\nu_T,\nu_Y)$ at each value of the treatment resistance U_T. This is the standard first-derivative LIV step and is useful here mainly because it prepares the second-derivative identification of the mediator effect itself.

[10] The error terms ν_Y and ϵ may be arbitrarily dependent. The distinction between ν_Y and ϵ allows the outcome to be stochastic even conditional on (T,M,ν_Y).

Proposition 4 (First-Derivative Identification of $\mathrm{E}(M(t) \mid U_T)$)**.** *Consider the structural model $T = f_T(Z, \nu_T)$, $M = f_M(T, \nu_T, \nu_Y)$ under: structural IV exogeneity $Z \perp\!\!\!\perp (\nu_T, \nu_Y, \epsilon)$; mediation exclusion (6) and consistency $M = M(T)$; separability (50) with $\phi(\nu_T)$ absolutely continuous (so that $U_T \sim \mathrm{Uniform}[0,1]$); and the regularity condition that $p \mapsto \mathrm{E}(M \cdot \mathbf{1}[T = t] \mid P_T(Z) = p)$ is differentiable for $t \in \{0,1\}$. Then*

$$\frac{\partial\, \mathrm{E}(M \cdot T \mid P_T(Z) = p)}{\partial p} = \mathrm{E}(M(1) \mid U_T = p) = \Pr\big(M(1) = 1 \mid U_T = p\big), \tag{52}$$

$$\frac{\partial\, \mathrm{E}(M \cdot (1-T) \mid P_T(Z) = p)}{\partial p} = -\,\mathrm{E}(M(0) \mid U_T = p) = -\Pr\big(M(0) = 1 \mid U_T = p\big). \tag{53}$$

The proof is in Appendix G.

Mediation Equation. We assume separability of the mediator equation in U_T and ν_Y:

$$M = \mathbf{1}[\xi(T, U_T) \geq \varphi(\nu_Y)], \tag{54}$$

where $\varphi(\nu_Y)$ is absolutely continuous and $\xi(t,u)$ is monotone in u for each t. We assume throughout that $\xi(t,\cdot)$ is *strictly decreasing* in u, so that individuals with lower U_T (stronger treatment propensity) have higher mediator propensity.[11] Separability (54) delivers the propensity score representation:

$$M = \mathbf{1}[\rho(T, U_T) \geq U_Y], \tag{55}$$

where $U_Y = F_{\varphi(\nu_Y)}(\varphi(\nu_Y)) \sim \mathrm{Uniform}[0,1]$ is the *mediator resistance*—higher values correspond to lower probability of $M = 1$ for any given $\rho(T, U_T)$, paralleling the treatment-resistance interpretation of U_T—and $\rho(T, U_T) = F_{\varphi(\nu_Y)}(\xi(T, U_T))$ is the *Marginal Mediator Response* (MMR). Since $U_T \perp\!\!\!\perp U_Y$ (by MCA),

$$\rho(t,u) \;=\; \Pr\left(M = 1 \mid T = t, U_T = u\right) \;=\; \mathrm{E}\left(M(t) \mid U_T = u\right)$$

for each (t,u). The MMR is the mediator analogue of the Marginal Treatment Response of Heckman and Vytlacil (2005): it gives the counterfactual probability that the mediator is active among individuals at treatment resistance $U_T = u$ when treatment is externally fixed at t, indexed by the latent treatment margin.

The structural object $\rho(t,p)$ differs from the observable within-stratum mediator rate

$$m_t(p) \equiv \Pr(M = 1 \mid T = t, P_T(Z) = p).$$

The object $m_t(p)$ averages mediator take-up over the truncated distribution of U_T induced by the treatment rule, whereas $\rho(t,p)$ is the mediator take-up probability exactly at the margin $U_T = p$. In particular,

$$m_1(p) = \Pr(M(1) = 1 \mid U_T \leq p) = \frac{1}{p}\int_0^p \rho(1,u)\, du, \tag{56}$$

$$m_0(p) = \Pr(M(0) = 1 \mid U_T > p) = \frac{1}{1-p}\int_p^1 \rho(0,u)\, du. \tag{57}$$

[11] The choice of decreasing (rather than increasing) is a normalization. The key requirement is strict monotonicity, which ensures a one-to-one mapping between U_T and the mediator resistance.

The structural object $\rho(t,p)$ is not an observed conditional probability; it is identified pointwise through the first derivative:

Corollary 2 (Identification of the Structural Mediator Propensity Score)**.** *Under the conditions of Proposition 4, MCA (14), and mediator separability (54):*

$$\rho(1,p) = \frac{\partial \,\mathrm{E}(M \cdot T \mid P_T(Z) = p)}{\partial p}, \tag{58}$$

$$\rho(0,p) = -\frac{\partial \,\mathrm{E}(M \cdot (1-T) \mid P_T(Z) = p)}{\partial p}. \tag{59}$$

The proof is in Appendix G.

Equivalently, the MMR admits the selection-corrected stratum-average form $\rho(1,p) = m_1(p) + p\,m_1'(p)$ and $\rho(0,p) = m_0(p) - (1-p)\,m_0'(p)$, expressing the take-up probability at the treatment margin as the observable within-stratum rate plus a selection correction.

Since $\xi(t,\cdot)$ is strictly decreasing and $F_{\varphi(\nu_Y)}$ is strictly increasing, $\rho(t,\cdot) = F_{\varphi(\nu_Y)}(\xi(t,\cdot))$ is *strictly decreasing* in U_T for each t. This is the **mediator monotonicity condition** in the LIV framework:

$$\rho(t,u) > \rho(t,u') \quad \text{for all } u < u' \text{ and each } t \in \{0,1\}. \tag{60}$$

Individuals with lower U_T (stronger treatment propensity) have a higher probability of $M(t) = 1$. The strict monotonicity ensures a one-to-one mapping between the propensity score p and the MMR $\rho(t,p)$, so the latter can serve as an alternative running variable for the LIV analysis. Since $\rho(t,p)$ is identified by Corollary 2, the monotonicity of $\rho(t,\cdot)$—and hence mediator separability (54) itself—is a *testable* shape restriction: a sign change in $\hat{\rho}'(t,p)$ across the support of $P_T(Z)$ is evidence against separability.

The outcome unobservable ν_Y appears in the mediator equation (54) because unobserved factors that affect Y also influence M—the source of endogeneity that necessitates instrumental variables. Under MCA, $U_T \perp\!\!\!\perp U_Y$, so compositional variation in $\rho(t,U_T)$ driven by Z through the propensity score is independent of the mediator resistance U_Y. This conditional exogeneity enables identification.

Identification of Counterfactual Outcomes. The central identification result exploits (49), which implies that the treatment resistance U_T is jointly independent of the counterfactual outcomes and the mediator resistance: $U_T \perp\!\!\!\perp (Y(t,m), U_Y)$. The resulting logic is sequential. The first derivative localizes the treatment margin at $U_T = p$. The second derivative maps that treatment margin into the mediator margin $U_Y = \rho(t,p)$. In consequence, the primitive mediator effect is identified from curvature, not slope. This is the main conceptual distinction between mediation LIV and the standard MTE framework.

Assumption 2 (Regularity Conditions for Second-Derivative Identification)**.**

(i) The propensity score $P_T(Z)$ has a continuously differentiable density on an open interval $(p_L, p_H) \subseteq (0,1)$.

(ii) The conditional expectations $p \mapsto \mathrm{E}(Y \cdot \mathbf{1}[M = m] \cdot \mathbf{1}[T = t] \mid P_T(Z) = p)$ and $p \mapsto \mathrm{E}(M \cdot \mathbf{1}[T = t] \mid P_T(Z) = p)$ are twice continuously differentiable on (p_L, p_H) for each $(t,m) \in \{0,1\}^2$.

(iii) *The mediator propensity score $\rho(t,\cdot)$ is continuously differentiable, with $\rho'(t,p) \neq 0$ for all $p \in (p_L, p_H)$. (Strict monotonicity of $\rho(t,\cdot)$ is already ensured by mediator separability (54) with $\xi(t,\cdot)$ strictly monotone; this condition adds smoothness.)*

(iv) Latent outcome regularity. *For each $(t,m) \in \{0,1\}^2$, the latent conditional mean function*

$$\mu_{tm}(u) \equiv \mathrm{E}\big(Y(t,m) \mid U_Y = u\big)$$

admits a bounded continuous version on the image $\rho(t,(p_L,p_H)) \subseteq (0,1)$. This rules out pathological discontinuities of the latent response curve at the chain-rule points $u = \rho(t,p)$ and is the latent counterpart of the observable smoothness condition in (ii).

Definition 1 (Marginal Mediator Effect)**.** *The* marginal mediator effect *at unobserved type $U_Y = u$ and treatment level t is:*

$$\Delta_M(t,u) \equiv \mathrm{E}(Y(t,1) - Y(t,0) \mid U_Y = u). \tag{61}$$

The MME is the nonparametric analogue of θ from the linear model. Like the marginal treatment effect of Heckman and Vytlacil (1999)—which is identified by the first derivative of $\mathrm{E}(Y \mid P_T(Z) = p)$—the MME is identified from derivatives of observable conditional expectations. The difference is that mediation requires one derivative to isolate the treatment margin and a second derivative to isolate the mediator margin induced by that treatment margin.

Theorem 2 (Second-Derivative Identification of Counterfactual Outcomes and the MME)**.** *Under conditions (6)–(8), MCA (14), consistency $M = M(T)$ and $Y = Y(T,M)$, separability conditions (50) and (54), and Assumption 2, for $(t,m) \in \{0,1\}^2$:*

$$\frac{\partial^2 \mathrm{E}\big(Y \cdot \mathbf{1}[M=m] \cdot \mathbf{1}[T=t] \,\big|\, P_T(Z)=p\big)}{\partial p^2} = (-1)^{1-m} \cdot (-1)^{1-t} \cdot \rho'(t,p) \cdot \mathrm{E}(Y(t,m) \mid U_Y = u), \tag{62}$$

where $u = \rho(t,p)$ is the MMR identified by Corollary 2. Define

$$G_t(p) \equiv \mathrm{E}\big(Y \cdot \mathbf{1}[T=t] \mid P_T(Z)=p\big), \qquad H_t(p) \equiv \mathrm{E}\big(M \cdot \mathbf{1}[T=t] \mid P_T(Z)=p\big). \tag{63}$$

Then the Marginal Mediator Effect satisfies

$$\Delta_M(t,u) = \frac{G_t''(p)}{H_t''(p)}, \qquad u = \rho(t,p). \tag{64}$$

Equivalently,

$$\Delta_M(t,u) = \frac{(-1)^{1-t}}{\rho'(t,p)} \cdot \frac{\partial^2 \mathrm{E}(Y \cdot \mathbf{1}[T=t] \mid P_T(Z)=p)}{\partial p^2}, \tag{65}$$

for p such that $\rho(t,p) = u$.

The proof is in Appendix G.

Equation (64) is the most transparent representation: the MME equals the curvature of the treatment-stratum outcome surface with respect to the treatment propensity score, divided by the curvature of the treatment-stratum mediator surface with respect to the same score. Equation (65) is the equivalent Jacobian form obtained from

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

Since $\rho'(t,p) \neq 0$ by Assumption 2(iii), both representations are well defined. The MME can also be written as a derivative with respect to the MMR $q_t(p) \equiv \rho(t,p)$; see Appendix G for details.[12]

The second-derivative result identifies the local primitives needed for nonlinear mediation analysis. Under full support of $P_T(Z)$ on $[0,1]$, Proposition 4 identifies the average counterfactual mediator,

$$\mathrm{E}[M(t)] \;=\; \int_0^1 \mathrm{E}[M(t) \mid U_T = u]\, du,$$

and Theorem 2 identifies the average counterfactual outcome,

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

With partial propensity-score support, $P_T(Z) \in (p_L, p_H) \subset (0,1)$, the MME is identified only on the image $\rho(t,(p_L,p_H))$. Aggregate parameters that integrate the MME are therefore point-identified only when their weighting functions place support inside that identified region, or when additional restrictions are imposed outside it, as in the standard MTE framework (Heckman and Vytlacil, 2005, Section 5).

The MME can be aggregated into population-level mediation parameters through weights that describe which mediator-resistance ranks are exposed to a counterfactual mediator state. For $t' \in \{0,1\}$, define the *mediator survival function*

$$S_M(t',u) \equiv \Pr\big(M(t') = 1 \mid U_Y = u\big).$$

Under the mediator LIV representation (55),

$$M(t') = \mathbf{1}\{\rho(t', U_T) \geq U_Y\}.$$

Hence, by MCA, $U_T \perp\!\!\!\perp U_Y$, and because $U_T \sim \mathrm{Uniform}[0,1]$,

$$S_M(t',u) = \Pr\big(\rho(t',U_T) \geq u \mid U_Y = u\big) = \int_0^1 \mathbf{1}\{\rho(t',p) \geq u\}\, dp.$$

With full support of $P_T(Z)$ on $[0,1]$, Corollary 2 identifies $\rho(t',p)$ for every $p \in [0,1]$, and therefore identifies $S_M(t',u)$. This survival function gives the aggregation formula for nested counterfactual outcomes. Since the mediator is binary, we can express $Y(t, M(t'))$ as following:

$$Y(t, M(t')) = Y(t,0) + \{Y(t,1) - Y(t,0)\}M(t').$$

[12] The two-stage nature of mediation carries a practical cost: second-derivative estimation is more demanding than the first-derivative estimation used in the standard MTE setting. This requires larger samples and stronger instrument variation than standard LIV.

Conditional on $U_Y = u$, the mediator state $M(t')$ depends on U_T, whereas $Y(t,1) - Y(t,0)$ depends on the outcome-side unobservables. MCA and the LIV representation therefore imply

$$\big(Y(t,0), Y(t,1)\big) \perp M(t') \mid U_Y.$$

It follows that

$$\mathrm{E}\big[\{Y(t,1) - Y(t,0)\}M(t') \mid U_Y = u\big] = \Delta_M(t,u)\, S_M(t',u).$$

Integrating over U_Y yields

$$\mathrm{E}[Y(t, M(t'))] = \mathrm{E}[Y(t,0)] + \int_0^1 \Delta_M(t,u) S_M(t',u)\, du.$$

The next proposition uses these survival weights to express population-level mediation parameters as integrals of the MME.

Proposition 5 (Mediation Parameters as Weighted Averages of the MME)**.** *Under the maintained assumptions of Theorem 2 and the denominator positivity* $0 < \mathrm{E}[M(t)] < 1$ *(so that* $\mathrm{MEM}(t)$ *and* $\mathrm{MEU}(t)$ *are well-defined), the following mediation parameters at treatment level* t *admit weighted-average representations in terms of* $\Delta_M(t,u)$*:*

(i) Average Mediator Effect *(analogue of ATE):*

$$\mathrm{AME}(t) \equiv \mathrm{E}[Y(t,1) - Y(t,0)] = \int_0^1 \Delta_M(t,u)\, du. \tag{66}$$

(ii) Mediator Effect on the Mediated *(analogue of TT):*

$$\mathrm{MEM}(t) \equiv \mathrm{E}[Y(t,1) - Y(t,0) \mid M(t) = 1] = \int_0^1 \Delta_M(t,u)\, \frac{S_M(t,u)}{\mathrm{E}(M(t))}\, du. \tag{67}$$

(iii) Mediator Effect on the Unmediated *(analogue of TUT):*

$$\mathrm{MEU}(t) \equiv \mathrm{E}[Y(t,1) - Y(t,0) \mid M(t) = 0] = \int_0^1 \Delta_M(t,u)\, \frac{1 - S_M(t,u)}{1 - \mathrm{E}(M(t))}\, du. \tag{68}$$

(iv) Natural Indirect Effect *(analogue of PRTE):*

$$\mathrm{NIE}(t) \equiv \mathrm{E}[Y(t, M(1)) - Y(t, M(0))] = \int_0^1 \Delta_M(t,u)\, \big[S_M(1,u) - S_M(0,u)\big]\, du. \tag{69}$$

(v) Natural Direct Effect*:*

$$\mathrm{NDE}(t) \equiv \mathrm{E}[Y(1, M(t)) - Y(0, M(t))] = \mathrm{E}[Y(1,0) - Y(0,0)] + \int_0^1 \{\Delta_M(1,u) - \Delta_M(0,u)\} S_M(t,u)\, du. \tag{70}$$

The signed weighting function $S_M(1,u) - S_M(0,u)$ *integrates to* $\mathrm{E}[M(1) - M(0)]$*; its sign at each* u *is unrestricted without mediator monotonicity. Under mediator monotonicity (*$M(1) \geq M(0)$*),* $S_M(1,u) -$

$S_M(0,u) \geq 0$ for all u, and the NIE (69) is a proper weighted average of $\Delta_M(t,u)$. Unlike the NIE, the NDE is not an aggregation of a single MME curve. It combines the baseline direct effect at $M=0$ with the treatment-state difference in the MME schedule, weighted by the counterfactual mediator survival function $S_M(t,u)$.

The proof is in Appendix G.

Remark 7 (Estimation)**.** *The ratio representation (64) suggests a direct estimation strategy. Estimate $G_t(p)$ and $H_t(p)$ by local polynomial regressions of $Y \cdot \mathbf{1}[T=t]$ and $M \cdot \mathbf{1}[T=t]$ on the propensity score $P_T(Z)$, using a polynomial order sufficient for second-derivative estimation, and compute*

$$\hat{\Delta}_M(t, \hat{\rho}(t,p)) = \frac{\hat{G}_t''(p)}{\hat{H}_t''(p)}.$$

This ratio avoids separate estimation of $\rho'(t,p)$, since $H_t''(p) = (-1)^{1-t}\rho'(t,p)$ absorbs the Jacobian. Standard results for local-polynomial derivative estimation apply (e.g., Fan and Gijbels, 1996); estimating second derivatives is more demanding than estimating the first derivatives used in the standard MTE setting. Inference for the ratio must account for estimation of the propensity score and for instability when $H_t''(p)$ is close to zero.

6.3 The Binary Instrument Case: Conditional Wald Ratios and Local Effects

This subsection specializes the analysis to the common case of a binary instrument, $Z \in \{z_0, z_1\}$. As in the LIV model of Section 6.2, the treatment $T \in \{0,1\}$ and the mediator $M \in \{0,1\}$ are also binary. The mediation IV conditions (6)–(8) and MCA (14) remain in force. Following Imbens and Angrist (1994), we assume:

Assumption 3 (Treatment Monotonicity)**.** *Treatment is weakly increasing in the instrument:*

$$T(z_1) \geq T(z_0) \qquad a.s. \tag{71}$$

Assumption 3 eliminates defiers: $\Pr(S_T = (1,0)) = 0$, and S_T takes values in $\{N, C, A\}$: never-takers N have $[T(z_0), T(z_1)]' = [0,0]'$, compliers C have $[0,1]'$, and always-takers A have $[1,1]'$. Under the mediation IV conditions and treatment monotonicity, the usual Wald ratios identify the standard LATE objects:[13]

$$\gamma^{LATE} \equiv \mathrm{E}[M(1) - M(0) \mid C] = \frac{\mathrm{E}[M \mid Z = z_1] - \mathrm{E}[M \mid Z = z_0]}{\mathrm{E}[T \mid Z = z_1] - \mathrm{E}[T \mid Z = z_0]}, \tag{72}$$

$$\psi^{LATE} \equiv \mathrm{E}[Y(1) - Y(0) \mid C] = \frac{\mathrm{E}[Y \mid Z = z_1] - \mathrm{E}[Y \mid Z = z_0]}{\mathrm{E}[T \mid Z = z_1] - \mathrm{E}[T \mid Z = z_0]}. \tag{73}$$

The parameter γ^{LATE} is the average mediator response among treatment compliers, while ψ^{LATE} combines the direct and indirect channels across the complier population.

[13] γ^{LATE} can be obtained from a 2SLS regression of M on T using Z as the instrument for T, and ψ^{LATE} from a 2SLS regression of Y on T using Z as the instrument for T.

Observable conditional Wald ratios. Conditioning on T changes the type composition within treatment strata. Under treatment monotonicity, the stratum $T = 0$ contains only never-takers and compliers, while the stratum $T = 1$ contains only always-takers and compliers. As Z moves from z_0 to z_1, the relative weights of these adjacent treatment types shift, generating within-stratum variation in the mediator through the collider mechanism developed in Section 6.1. This motivates the conditional Wald ratio[14]

$$\theta^W(t) \equiv \frac{\operatorname{cov}(Y, Z \mid T = t)}{\operatorname{cov}(M, Z \mid T = t)} = \frac{\mathrm{E}[Y \mid Z = z_1, T = t] - \mathrm{E}[Y \mid Z = z_0, T = t]}{\mathrm{E}[M \mid Z = z_1, T = t] - \mathrm{E}[M \mid Z = z_0, T = t]}, \qquad t \in \{0, 1\}. \tag{74}$$

Under MCA and treatment monotonicity, $\theta^W(t)$ is well defined whenever the conditional first stage is nonzero, but its *causal interpretation* as a local mediator effect requires an additional monotonicity condition on the adjacent treatment-type margins that generate the first stage.

Compositional mediator monotonicity. The binary conditional first stage does not compare $M(1)$ to $M(0)$ globally. Instead, it compares mediator take-up across adjacent treatment-type blocks within each treatment stratum: N versus C inside $T = 0$, and C versus A inside $T = 1$. The monotonicity condition must therefore target those margins.

Under the structural mediator equation in Section 6.2, define the type-indexed mediator potential

$$M(t, s) \equiv \mathbf{1}\left[\rho(t, U_T^{(s)}) \geq U_Y\right], \tag{75}$$

where $U_T^{(s)}$ denotes U_T on the resistance block corresponding to treatment type s (e.g., $U_T^{(N)} \in (p_1, 1]$, $U_T^{(C)} \in (p_0, p_1]$, $U_T^{(A)} \in [0, p_0)$). This object evaluates mediator take-up at a fixed treatment level t while moving across treatment-type blocks and holding the mediator resistance U_Y fixed.

Assumption 4 (Compositional Mediator Monotonicity)**.** *The adjacent type margins generating the conditional first stages contain no mediator defiers:*

$$M(1, A) \geq M(1, C) \qquad \textit{a.s.}, \tag{76}$$

$$M(0, C) \geq M(0, N) \qquad \textit{a.s.} \tag{77}$$

Assumption 4 states that, at a fixed treatment level, mediator take-up is ordered by treatment propensity: always-takers exceed compliers, and compliers exceed never-takers.[15] Intuitively, the assumption says that the latent disposition driving selection into treatment also drives uptake of the mediator: among individuals held at a common treatment level, those with stronger treatment propensity (the always-takers within $T = 1$, the compliers within $T = 0$) are weakly more likely to activate the mediator. Just as treatment monotonicity orders individuals by a single resistance to treatment, compositional monotonicity requires that this same ordering govern mediator take-up across adjacent treatment-type blocks.

With a binary instrument, this adjacent-block restriction is not directly testable because $\rho(t, \cdot)$ is not identified pointwise. Section 6.4 instead develops Kitagawa-style tests of the joint implications of MCA

[14] Operationally, $\theta^W(t)$ is the just-identified 2SLS coefficient from the subsample $T = t$ regression of Y on M using Z as the instrument for M. Equivalently, estimate the pooled saturated IV specification $Y = \alpha + \tau T + \theta M + \rho(MT) + e$, treating M and MT as endogenous and using Z and ZT as excluded instruments, with $(1, T)$ included as exogenous regressors. Then $\theta = \theta^W(0)$ and $\theta + \rho = \theta^W(1)$.

[15] Compositional Mediator Monotonicity is equivalently to state that $\Pr\big(M(0, N) = 1,\ M(0, C) = 0\big) = 0$, and $\Pr\big(M(1, C) = 1,\ M(1, A) = 0\big) = 0$.

and compositional mediator monotonicity.

Remark 8 (Monotonicities under the LIV structure)**.** *The monotonicity conditions used in the binary-instrument model are inherited from the LIV structure. After orienting the instrument so that $P_T(z_1) \geq P_T(z_0)$, treatment separability* (50) *implies Treatment Monotonicity (Assumption 3). Likewise, mediator separability* (54) *implies Compositional Mediator Monotonicity (Assumption 4), because $\rho(t, \cdot)$ is strictly decreasing in treatment resistance and the adjacent treatment-type blocks are ordered in U_T.*

Remark 9 (Cross-treatment mediator monotonicity)**.** *Navjeevan, Pinto and Santos (2026) impose the cross-treatment restriction $M(1) \geq M(0)$ a.s. — the direct mediator analogue of treatment monotonicity, requiring that the treatment weakly raises mediator uptake at the individual level. Compositional Mediator Monotonicity (Assumption 4) is logically distinct: it orders mediator propensity* across types *within each treatment state ($A \succeq C \succeq N$) rather than* across treatment states *within each individual. Cross-treatment is a causal restriction on $T \to M$; compositional is a selection restriction on the joint distribution of treatment and mediator propensities.*

The two support different identifications. Cross-treatment monotonicity, combined with the auxiliary exogeneity restrictions of Navjeevan, Pinto and Santos (2026), delivers direct treatment effects. Compositional monotonicity gives the conditional Wald ratios in (74) *a local-mediator-effect interpretation. The LIV threshold representation implies compositional monotonicity automatically (Remark 8). The cross-treatment monotonicity is a stronger assumption which requires the additional ordering $\rho(1, u) \geq \rho(0, u)$. Appendix G.2 compares the resulting type structures.*

Local mediator effects. The conditional Wald ratios isolate the mediator's outcome effect for a specific subpopulation: individuals whose mediator status switches across the adjacent treatment-type margin that drives the within-stratum first stage. We call this subpopulation *compositional mediator compliers at $T = t$*, in direct analogy with treatment compliers from Assumption 3. Just as treatment compliers are those whose treatment status switches with the instrument, mediator compliers at $T = t$ are those whose mediator status switches across the adjacent type-composition margin within the $T = t$ stratum.

Formally, define the compositional mediator-complier events

$$\mathcal{C}_0^M \equiv \{M(0, C) > M(0, N)\}, \tag{78}$$
$$\mathcal{C}_1^M \equiv \{M(1, A) > M(1, C)\}, \tag{79}$$

and the corresponding local mediator effects

$$\theta^{LM}(0) \equiv \mathrm{E}\left[Y(0, 1) - Y(0, 0) \mid \mathcal{C}_0^M\right], \tag{80}$$
$$\theta^{LM}(1) \equiv \mathrm{E}\left[Y(1, 1) - Y(1, 0) \mid \mathcal{C}_1^M\right]. \tag{81}$$

The parameter $\theta^{LM}(t)$ is the average outcome effect of switching the mediator from 0 to 1 among the mediator compliers at $T = t$. It is the finite-difference, binary-instrument analogue of the pointwise Marginal Mediator Effect $\Delta_M(t, u)$ from Section 6.2, averaged over the local compositional margin induced by the instrument.

Lemma 1 (Conditional exogeneity across treatment types)**.** *Under MCA* (14) *and IV exogeneity $Z \perp\!\!\!\perp$*

(ν_T, ν_Y, ϵ),

$$Y(t,m) \perp\!\!\!\perp S_T \mid T \qquad \textit{for all } (t,m).$$

The proof is in Appendix G.

Lemma 1 is the conditional-exogeneity step behind Theorem 3. The conditional Wald ratio $\theta^W(t)$ is generated by the instrument reshuffling treatment-type composition within the stratum $T = t$; the lemma guarantees that this reshuffling does not also reshuffle the distribution of potential outcomes. Thus, conditional on $T = t$, the induced change in mediator take-up is interpretable as mediator-channel variation rather than confounding variation in the potential outcomes.

Theorem 3 (Conditional Wald identification of local mediator effects)**.** *Suppose that* $\Pr(S_T = s) > 0; s \in \{N, C, A\}$, *and* $\Pr\left(\mathcal{C}_t^M\right) > 0; t \in \{0, 1\}$. *Under conditions* (6)–(8)*, MCA* (14)*, and Monotonicity Assumptions 3 and 4, the conditional Wald ratios identify the local mediator effects:*

$$\theta^W(0) = \mathrm{E}\left[Y(0,1) - Y(0,0) \mid \mathcal{C}_0^M\right] = \theta^{LM}(0), \tag{82}$$

$$\theta^W(1) = \mathrm{E}\left[Y(1,1) - Y(1,0) \mid \mathcal{C}_1^M\right] = \theta^{LM}(1). \tag{83}$$

The proof is in Appendix G.

Theorem 3 shows that the binary-instrument design identifies a genuinely local mediator effect. It evaluates the mediator effect for the individuals whose mediator status is shifted by the adjacent treatment-type margin operating within a fixed treatment state: the C-versus-N margin for $t = 0$, and the A-versus-C margin for $t = 1$. Proposition 6 expresses $\theta^{LM}(t)$ as a weighted average of the Marginal Mediator Effect $\Delta_M(t,u)$, where the weights describe which mediator-resistance ranks are most affected by the relevant compositional shift.

Proposition 6 (The Local Mediator Effect as a Weighted Average of the MME)**.** *Under the conditions of Theorem 3, the local mediator effect* $\theta^{LM}(t)$ *admits a weighted-average representation in terms of the Marginal Mediator Effect* $\Delta_M(t,u)$ *(Definition 1):*

$$\theta^{LM}(t) = \frac{\displaystyle\int_0^1 \Delta_M(t,v)\,\omega_t(v)\,dv}{\displaystyle\int_0^1 \omega_t(v)\,dv}, \tag{84}$$

where the weight function is the difference in survival functions of the MMR across adjacent treatment-type blocks:

$$\omega_t(v) \equiv S_{s_H}(v) - S_{s_L}(v), \qquad S_s(v) \equiv \Pr\big(\rho(t,U_T) \geq v \;\big|\; S_T = s\big). \tag{85}$$

For $t = 0$*:* $s_H = C$ *and* $s_L = N$*. For* $t = 1$*:* $s_H = A$ *and* $s_L = C$*. Under compositional mediator monotonicity,* $\omega_t(v) \geq 0$ *for all* v*, so* (84) *is a proper weighted average. The normalizing constant* $\int_0^1 \omega_t(v)\,dv = \mathrm{E}[\rho(t,U_T) \mid s_H] - \mathrm{E}[\rho(t,U_T) \mid s_L]$ *equals the type-conditional first stage—the difference in mean mediator response across adjacent treatment-type blocks. As the propensity-score gap* $p_1 - p_0$ *shrinks, both survival functions converge to step functions, the weights concentrate at a point, and* $\theta^{LM}(t)$ *converges to the pointwise MME* $\Delta_M(t, \rho(t,p))$ *of Theorem 2.*

The proof is in Appendix G.

The weight $\omega_t(v)$ measures how much the compositional shift from type s_L to type s_H increases the probability of $M = 1$ at mediator resistance $U_Y = v$. Individuals at U_Y values where the compositional shift has a large effect on mediator status receive more weight. This is the mediation analogue of the Heckman–Vytlacil weighting scheme, in which different treatment parameters weight the MTE differently depending on the source of identifying variation. Proposition 6 is the precise bridge between the binary-instrument and continuous-support models: the conditional Wald identifies $\theta^{LM}(t)$, and the weighted-average representation shows that this estimand is a local average of the same MME primitive that the LIV model identifies pointwise.

Remark 10 (Connection to the linear model)**.** *If mediator effects are constant, $Y_i(t,1) - Y_i(t,0) = \theta(t)$ for all i, then $\Delta_M(t,u) = \theta(t)$ for all u, so $\theta^{LM}(t) = \theta(t) = \text{AME}(t)$ and the conditional Wald reduces to the estimand in Section 4, giving $\text{NIE}_C(t) = \gamma^{LATE} \cdot \theta(t) = \theta \cdot \gamma$.*

Remark 11 (Estimation)**.** *The binary case is estimated with three Wald-type ratios:*

(i) Treatment–mediator effect*:*

$$\hat{\gamma} = \frac{\bar{M}_{Z=z_1} - \bar{M}_{Z=z_0}}{\bar{T}_{Z=z_1} - \bar{T}_{Z=z_0}}.$$

(ii) Total effect*:*

$$\hat{\psi} = \frac{\bar{Y}_{Z=z_1} - \bar{Y}_{Z=z_0}}{\bar{T}_{Z=z_1} - \bar{T}_{Z=z_0}}.$$

(iii) Local mediator effect*: for each $t \in \{0,1\}$,*

$$\hat{\theta}^{LM}(t) = \frac{\bar{Y}_{Z=z_1,T=t} - \bar{Y}_{Z=z_0,T=t}}{\bar{M}_{Z=z_1,T=t} - \bar{M}_{Z=z_0,T=t}}.$$

Under Theorem 3, $\hat{\theta}^{LM}(t)$ estimates the local mediator effect within treatment stratum t. The decomposition $\psi^{LATE} = NDE_C(t) + NIE_C(1-t)$ admits two versions: at reference level $t = 0$, the indirect component is $\hat{\gamma}\,\hat{\theta}^{LM}(1)$ with residual $\hat{\psi} - \hat{\gamma}\,\hat{\theta}^{LM}(1)$; at reference level $t = 1$, it is $\hat{\gamma}\,\hat{\theta}^{LM}(0)$ with residual $\hat{\psi} - \hat{\gamma}\,\hat{\theta}^{LM}(0)$. Under constant mediator effects the two coincide. Under heterogeneity, both should be reported. Standard errors may be obtained by the delta method or by the nonparametric bootstrap.

Because the conditional Wald ratios are estimated within the subsamples $T = t$, weak-instrument concerns are more severe than in the unconditional first stage. Applied work should therefore report the conditional first-stage strength within each treatment stratum and rely on Anderson–Rubin-style inference when the conditional first stage is weak.

6.4 Kitagawa-Style Falsification Tests

We now turn from identification to falsification. The binary-instrument mediation model implies a set of distributional inequalities in the spirit of Kitagawa (2015) and Kwon and Roth (2026). In the Kwon and Roth setting, treatment is randomized and the triple (Y, M, T) forms a standard LATE structure for the mediator. In our setting, the relevant object is instead the conditional triple $(Y, M, Z) \mid T = t$: once one conditions on treatment status, MCA supplies conditional exogeneity and compositional variation supplies conditional relevance. The resulting inequalities should be interpreted as *falsification* tests for the mediation-specific identifying package, not as validation of that package.

The tests are organized as a hierarchy: (i) two "baseline" tests that probe the standard LATE package for $Z \to T$ using M or Y as the outcome (these do not require MCA); (ii) a conditional test that exploits the conditional-IV structure for M within treatment strata enabled by MCA.

Baseline tests. Kitagawa (2015) shows that instrument validity and treatment monotonicity ($T(z_1) \geq T(z_0)$) jointly imply, for any outcome W and all Borel sets $A \subseteq \mathbb{R}$:

$$\Pr(W \in A,\, T = 0 \mid Z = z_0) \;\geq\; \Pr(W \in A,\, T = 0 \mid Z = z_1), \tag{86}$$

$$\Pr(W \in A,\, T = 1 \mid Z = z_1) \;\geq\; \Pr(W \in A,\, T = 1 \mid Z = z_0). \tag{87}$$

We apply these inequalities twice: first with $W = M$ (Test 1, probing the LATE package for the causal effect of T on M) and then with $W = Y$ (Test 2, probing the LATE package for the total causal effect of T on Y). Passing both tests supports the maintained LATE structure for $Z \to T$; importantly, they do not validate mediation.

Conditional Kitagawa test for MCA. MCA adds a distinctive implication: conditional on realized treatment status $T = t$, the instrument Z is exogenous for $Y(t, m)$ and relevant for M via compositional variation (Section 6.1). Compositional mediator monotonicity (Assumption 4) provides the no-defier restriction needed to apply the Kitagawa (2015) logic to the conditional-IV model $(Y, M, Z) \mid T = t$.

The argument parallels the standard Kitagawa construction. Within $T = 1$, changing Z from z_0 to z_1 adds treatment compliers (C) to the stratum alongside always-takers (A). Assumption 4 gives $M(1, A) \geq M(1, C)$ almost surely: at every value of the latent mediator resistance outside a null set, the mediator under the A-type propensity is weakly higher than under the C-type propensity. This yields the set inclusion $\{M(1, C) = 1\} \subseteq \{M(1, A) = 1\}$ almost surely—the "no defiers" condition that Kitagawa (2015) requires. In finite-unit language, no unit has $M = 1$ under the C-type propensity but $M = 0$ under the A-type propensity. Combined with conditional exogeneity ($Y(t, m) \perp\!\!\!\perp S_T \mid T$, from MCA and IV-core), the distributional inequalities follow for all Borel sets $A \subseteq \mathbb{R}$ and each $t \in \{0, 1\}$:

$$\Pr(Y \in A,\, M = 0 \mid Z = z_0, T = t) \;\geq\; \Pr(Y \in A,\, M = 0 \mid Z = z_1, T = t), \tag{88}$$

$$\Pr(Y \in A,\, M = 1 \mid Z = z_1, T = t) \;\geq\; \Pr(Y \in A,\, M = 1 \mid Z = z_0, T = t), \tag{89}$$

after orienting (z_0, z_1) within each stratum so that $\Delta_t \equiv \Pr(M = 1 \mid Z = z_1, T = t) - \Pr(M = 1 \mid Z = z_0, T = t) \geq 0$.[16]

Interpretation. The baseline tests (86)–(87) probe the standard LATE package for $Z \to T$. The conditional inequalities (88)–(89) probe the additional mediation-specific restrictions: MCA, which delivers conditional exogeneity, and compositional mediator monotonicity, which delivers the relevant conditional monotonicity. Hence rejection after the baseline tests are passed isolates failure of the *mediation-specific package*, but it does not distinguish whether the failure comes from MCA or from compositional monotonicity. Implementation of the conditional test via moment inequalities, including the test statistic $\widehat{T}_{\text{MCA}}$ and bootstrap critical values, is detailed in Appendix G.3.

[16] The orientation may differ from the treatment equation. Within $T = 0$, compositional monotonicity ($M(0, C) \geq M(0, N)$) implies that removing compliers (moving from $Z = z_0$ to $Z = z_1$) weakly reduces $\Pr(M = 1 \mid T = 0)$, so $\Delta_0 \leq 0$ under the treatment-equation orientation. The sign normalization swaps (z_0, z_1) within the $T = 0$ stratum to obtain $\Delta_0 \geq 0$.

7 Simulation Evidence

This section evaluates the finite-sample performance of the mediation estimator developed in Section 4 and the diagnostic tools of Section 5. We design a data-generating process that satisfies the structural equations (21)–(23) exactly, with a transparent Gaussian error structure that allows systematic variation of the key model parameters. Four experiments examine: (i) identification and estimation under correct specification, (ii) the relationship between instrument strength and conditional first-stage power, (iii) calibration of the κ-based and R^2-based sensitivity analyses, and (iv) size and power of the specification tests for mediated confounding. Experiments 1–3 use a single instrument ($K = 1$), the standard setting for the mediation decomposition; Experiment 4 introduces a second instrument with a direct effect on M to evaluate the overidentification tests of Section 5.1. We then compare the one-instrument estimator against the two-instrument benchmark of Frölich and Huber (2017a) under correct specification, and verify robustness to non-Gaussian error distributions.

7.1 Data-Generating Process

The structural equations follow the linear mediation model of Section 4.1, with all variables orthogonalized with respect to $[\mathbf{1}, X]$:

$$T_i = \pi_1 Z_{1i} + \pi_2 Z_{2i} + \varepsilon_{i,T}, \tag{90}$$

$$M_i = \gamma \cdot T_i + \varepsilon_{i,M}, \tag{91}$$

$$Y_i = \tau \cdot T_i + \theta \cdot M_i + \beta_X X_i + \varepsilon_{i,Y}, \tag{92}$$

where $Z_{1i}, Z_{2i}, X_i \overset{iid}{\sim} \mathcal{N}(0,1)$ are mutually independent and independent of the structural errors. When $\pi_2 = 0$, the model reduces to the just-identified case ($K = 1$); when $\pi_2 \neq 0$, the additional instrument provides variation in M independent of the collider channel through T, enabling overidentification tests. The covariate X is included for sensitivity calibration; when $\beta_X = 0$, it plays no role.

The structural errors follow a joint normal distribution:

$$(\varepsilon_{i,T},\, \varepsilon_{i,M},\, \varepsilon_{i,Y})' \sim \mathcal{N}\left(\mathbf{0},\, \begin{pmatrix} 1 & \rho_{TM} & \rho_{TY} \\ \rho_{TM} & 1 & \rho_{MY} \\ \rho_{TY} & \rho_{MY} & 1 \end{pmatrix}\right), \tag{93}$$

where ρ_{TM}, ρ_{MY}, and ρ_{TY} govern treatment–mediator, mediator–outcome, and treatment–outcome confounding, respectively. Positive definiteness requires $1 - \rho_{TM}^2 - \rho_{MY}^2 - \rho_{TY}^2 + 2\rho_{TM}\rho_{MY}\rho_{TY} > 0$.

The baseline calibration is:

$$\pi_1 = 0.5, \quad \gamma = 0.4, \quad \tau = 0.3, \quad \theta = 0.5, \quad \rho_{TM} = 0.5, \quad \rho_{MY} = 0.3, \quad \rho_{TY} = 0. \tag{94}$$

This configuration generates a strong first stage for T and a weaker conditional first stage for M, reflecting the fact that the identifying variation for the mediator operates through the collider mechanism rather than a direct instrument channel.

7.2 Experiment 1: Finite-Sample Performance

We first verify that the 2SLS estimator recovers all four mediation parameters with correct bias, root mean squared error (RMSE), and confidence interval coverage when MCA holds. The design fixes $\rho_{TY} = 0$ (MCA satisfied), $\rho_{MY} = 0.3$, and $\rho_{TM} = 0.5$, varying sample size $n \in \{500, 1{,}000, 2{,}500, 5{,}000\}$ and instrument strength $\pi_1 \in \{0.2, 0.5, 1.0\}$. We use $K = 1$ instrument ($\pi_2 = 0$, $\beta_X = 0$) and conduct $R = 2{,}000$ Monte Carlo replications for each configuration.

Table 3: Experiment 1: Finite-Sample Performance of Mediation Parameters

		Mediator Effect $\hat{\theta}$ (True = 0.5)				Direct Effect $\hat{\tau}$ (True = 0.3)			
n	π_1	Bias	RMSE	Cov.	SE	Bias	RMSE	Cov.	SE
Panel A: MCA holds ($\rho_{TY} = 0$)									
500	0.2	−0.099	2.089	0.974	3.045	0.087	1.873	0.970	2.718
1,000	0.2	−0.037	0.529	0.973	0.476	0.032	0.461	0.973	0.418
2,500	0.2	−0.012	0.214	0.966	0.216	0.010	0.189	0.965	0.191
5,000	0.2	−0.010	0.148	0.962	0.149	0.010	0.131	0.962	0.132
500	0.5	−0.014	0.208	0.970	0.212	0.011	0.171	0.969	0.174
1,000	0.5	−0.006	0.144	0.959	0.145	0.004	0.118	0.965	0.120
2,500	0.5	−0.002	0.089	0.954	0.090	0.001	0.073	0.954	0.075
5,000	0.5	−0.002	0.063	0.955	0.064	0.001	0.051	0.960	0.053
500	1.0	−0.005	0.130	0.955	0.129	0.003	0.090	0.947	0.090
1,000	1.0	−0.004	0.089	0.959	0.091	0.003	0.061	0.965	0.063
2,500	1.0	−0.002	0.056	0.953	0.057	0.001	0.039	0.952	0.040
5,000	1.0	0.000	0.041	0.942	0.040	−0.001	0.029	0.949	0.028
Panel B: MCA violated ($n = 2{,}500$, $\pi_1 = 0.5$)									
2,500	0.5[a]	0.198	0.216	0.366	0.087	−0.079	0.106	0.794	0.072
2,500	0.5[b]	0.601	0.608	0.000	0.090	−0.241	0.252	0.080	0.074

Notes: 2,000 replications. Panel A: $\rho_{TY} = 0$ (MCA holds). Panel B: [a]$\rho_{TY} = 0.1$; [b]$\rho_{TY} = 0.3$. Baseline: $\gamma = 0.4$, $\rho_{TM} = 0.5$, $\rho_{MY} = 0.3$. Cov. = empirical 95% CI coverage rate. SE = average estimated standard error.

We report bias, RMSE, 95% confidence interval coverage, and estimated standard errors for each parameter. Table 3 presents results for the mediation parameters $(\hat{\theta}, \hat{\tau})$.[17]

Results. Under correct specification ($\rho_{TY} = 0$), the estimators for θ and τ are essentially unbiased across all configurations. Coverage is close to the nominal 95% level when the instrument is moderate or strong ($\pi_1 \in \{0.5, 1.0\}$). When $\pi_1 = 0.2$, the conditional first-stage is weak, leading to inflated standard errors and overcoverage for $\hat{\theta}$, with correspondingly large RMSE. As expected, RMSE decreases at the $1/\sqrt{n}$ rate across all configurations.

Behavior under MCA violation. The bottom panel of Table 3 reports results for $\rho_{TY} \in \{0.1, 0.3\}$ at the baseline calibration ($n = 2{,}500$, $\pi_1 = 0.5$). Violations induce bias in $\hat{\theta}$ (positive) and $\hat{\tau}$ (negative), with opposite signs maintaining the total effect decomposition. At $\rho_{TY} = 0.3$, the bias is substantial and coverage deteriorates sharply. These results highlight the importance of the diagnostic tools developed in Section 5.

[17] Results for the standard IV components $(\hat{\pi}_1, \hat{\gamma}, \hat{\tau}^{\text{total}})$ are essentially unbiased with near-nominal coverage across configurations and are omitted to conserve space.

7.3 Experiment 2: Conditional First-Stage Diagnostics

Remark 2 shows that the conditional first-stage F-statistic $F_{MZ|T}$ depends on $\rho_{TM}^2 r_{TZ}^2/(1-\rho_{TM}^2)$, whereas the standard first-stage F-statistic F_{TZ} depends only on $r_{TZ}^2/(1-r_{TZ}^2)$. Thus, a strong standard first stage does not guarantee a strong conditional first stage. This experiment validates this prediction and documents its finite-sample implications.

The design fixes $n = 2{,}500$, $\pi_1 = 0.5$, $\rho_{MY} = 0.3$, $\rho_{TY} = 0$, and varies $\rho_{TM} \in \{0.1, 0.2, 0.3, 0.5, 0.8\}$, with $R = 2{,}000$ replications.

Results. Table 4 confirms the theoretical predictions. The standard first-stage F-statistic (F_{TZ}) is invariant to ρ_{TM}, while the conditional first-stage F-statistic ($F_{MZ|T}$) increases monotonically with $|\rho_{TM}|$, closely matching the theoretical values.

The finite-sample implications are substantial. When $\rho_{TM} = 0.1$, the conditional first stage is weak ($F_{MZ|T} \approx 5$), leading to large RMSE and severely distorted inference for $\hat{\theta}$. As ρ_{TM} increases, estimator performance improves monotonically. The transition to reliable inference occurs around $\rho_{TM} \approx 0.3$ (conditional $F \approx 49$), consistent with conventional thresholds. Notably, F_{TZ} remains large throughout, so relying on the standard first stage alone would mask this fragility.

Table 4: Experiment 2: Conditional First-Stage Diagnostics

	First-Stage F-statistics			Mediator Effect $\hat{\theta}$ (True = 0.5)				
ρ_{TM}	F_{TZ}	$F_{MZ\|T}$	$F_{MZ\|T}^{\text{theory}}$	Bias	RMSE	Cov.	Cov.$_{\text{AR}}$	SE
0.1	622.6	5.1 [2.5, 8.6]	5.0	−0.063	6.185	0.985	0.953	26.889
0.2	624.9	20.9 [15.2, 27.7]	20.8	−0.018	0.249	0.969	0.956	0.246
0.3	627.0	49.5 [40.5, 59.5]	49.4	−0.002	0.157	0.953	0.941	0.154
0.5	625.5	167.5 [150.3, 184.5]	166.5	−0.003	0.090	0.958	0.955	0.090
0.8	624.4	887.4 [840.4, 933.7]	887.8	0.001	0.057	0.944	0.946	0.056

Notes: 2,000 replications, $n = 2{,}500$, $\pi_1 = 0.5$, $\rho_{MY} = 0.3$, $\rho_{TY} = 0$. F_{TZ} = F-statistic for $T \sim Z$. $F_{MZ|T}$ = F-statistic for Z in $M \sim Z + T$ (i.e., $t^2_{Z|T}$); reported as median [IQR]. $F_{MZ|T}^{\text{theory}}$ = theoretical prediction from Remark 2. Cov. = conventional 2SLS 95% CI coverage. Cov.$_{\text{AR}}$ = Anderson–Rubin 95% CI coverage.

7.4 Experiment 3: Sensitivity Analysis Calibration

Section 5.2 develops two complementary sensitivity frameworks: the R^2-based approach adapted from Cinelli and Hazlett (2024), which quantifies robustness of the conditional reduced form to omitted confounders, and the κ-based approach (Section 5.2.1), which directly targets the MCA assumption by working with $\kappa = \text{Cov}(\varepsilon_T, \varepsilon_Y)$. This experiment evaluates both approaches.

The design uses $n = 2{,}500$, $\pi_1 = 0.5$ ($K = 1$), $\rho_{TM} = 0.5$, $\rho_{MY} = 0.3$, and an observed covariate X with $\beta_X = 0.3$. We conduct $R = 2{,}000$ replications at each value of $\rho_{TY} \in \{0, 0.10, 0.20, 0.30, 0.50\}$.

κ-Based Sensitivity. Table 5 reports the breakeven values $\hat{\kappa}^*$ and $\hat{\kappa}^*_{\text{CI}}$. Under correct specification ($\rho_{TY} = 0$), the median breakeven is $\hat{\kappa}^* = 0.249$, while the CI threshold is $\hat{\kappa}^*_{\text{CI}} = 0.160$. These values quantify the magnitude of violation required to overturn the estimate.

κ-Adjusted Coverage. The final column of Table 5 reports coverage of the κ-adjusted confidence interval (46) evaluated at the true violation $\kappa_0 = \rho_{TY}$. Coverage remains close to nominal levels (0.938–0.975) across all configurations, confirming that the adjustment recovers correct inference when the violation magnitude is known.

Limitation of R^2-Based Robustness Values. The R^2-based robustness value increases with ρ_{TY}, from 0.008 to 0.103, implying that larger violations appear more robust. This occurs because MCA violations bias $\hat{\theta}$ away from zero, making it mechanically harder for reduced-form confounders to overturn the estimate. The R^2-based framework therefore does not capture violations of MCA, which operate through the ε_T–ε_Y channel rather than the $\tilde{Z}$–$\tilde{Y}$ relationship.

Complementarity. The two frameworks address different threats. The κ-based approach directly assesses violations of MCA and is applicable in the $K = 1$ setting, while the R^2-based approach captures reduced-form confounding. Both should be reported, but only the κ-based analysis is informative about the identifying assumption.

Table 5: Experiment 3: Sensitivity Analysis Calibration

		κ-Based Sensitivity			Conventional		R^2-Based
ρ_{TY}	κ	Med. $\hat{\kappa}^*$	Med. $\hat{\kappa}^*_{\text{CI}}$	Cov.$_\kappa$	Bias($\hat{\theta}$)	Cov.	Med. RV
0	0	0.249	0.160	0.958	−0.003	0.958	0.008
0.10	0.10	0.351	0.266	0.946	0.200	0.363	0.017
0.20	0.20	0.449	0.365	0.938	0.402	0.002	0.029
0.30	0.30	0.550	0.462	0.957	0.601	0.000	0.046
0.50	0.50	0.752	0.647	0.975	1.007	0.000	0.103

Notes: 2,000 replications with $n = 2{,}500$, $\pi_1 = 0.5$, $\rho_{TM} = 0.5$, $\beta_X = 0.3$. $\kappa = \text{Cov}(\varepsilon_T, \varepsilon_Y) = \rho_{TY}$ (unit-variance errors). Med. $\hat{\kappa}^*$ = median breakeven (43): the minimum κ needed to reduce $\hat{\theta}$ to zero. Med. $\hat{\kappa}^*_{\text{CI}}$ = median CI breakeven (44): the minimum κ to bring the 95% CI to include zero. Cov.$_\kappa$ = coverage of the κ-adjusted CI (46) at the true $\kappa_0 = \rho_{TY}$. Cov. = conventional 95% CI coverage. Med. RV = median $\text{RV}_{1,0.05}(\hat{\theta})$ from the R^2-based framework; RV increases with ρ_{TY} because the biased estimate is further from zero, making it mechanically harder to overturn (Appendix F).

Visual Summary. Figure 1 plots the bias in $\hat{\theta}$ and the 95% CI coverage rate as ρ_{TY} increases. Bias increases approximately linearly, while coverage deteriorates rapidly, falling below 40% by $\rho_{TY} = 0.1$. The breakeven values in Table 5 provide a direct way to assess whether such violations are plausible in practice.

7.5 Experiment 4: Specification Test Size and Power

We evaluate the size and power of the J-test (139), the t-test (140) targeting the MCA moment, and the likelihood-ratio (LR) test (40) when additional instruments are available ($K \geq 2$).

Consistent with the rank-condition discussion in Section 5.1, test power requires instruments that provide variation in M beyond the indirect channel through T. The design uses $K = 2$ instruments with $\pi_1 = 0.5$ and $\pi_2 = 0.3$, and augments the mediator equation with a direct instrument effect $\phi = 0.3$, yielding one degree of overidentification from the MCA moment. We fix $n = 2{,}500$ and $R = 2{,}000$ replications. To assess size, we set $\rho_{TY} = 0$; to assess power, we vary $\rho_{TY} \in \{0, 0.10, 0.20, 0.30, 0.50\}$.

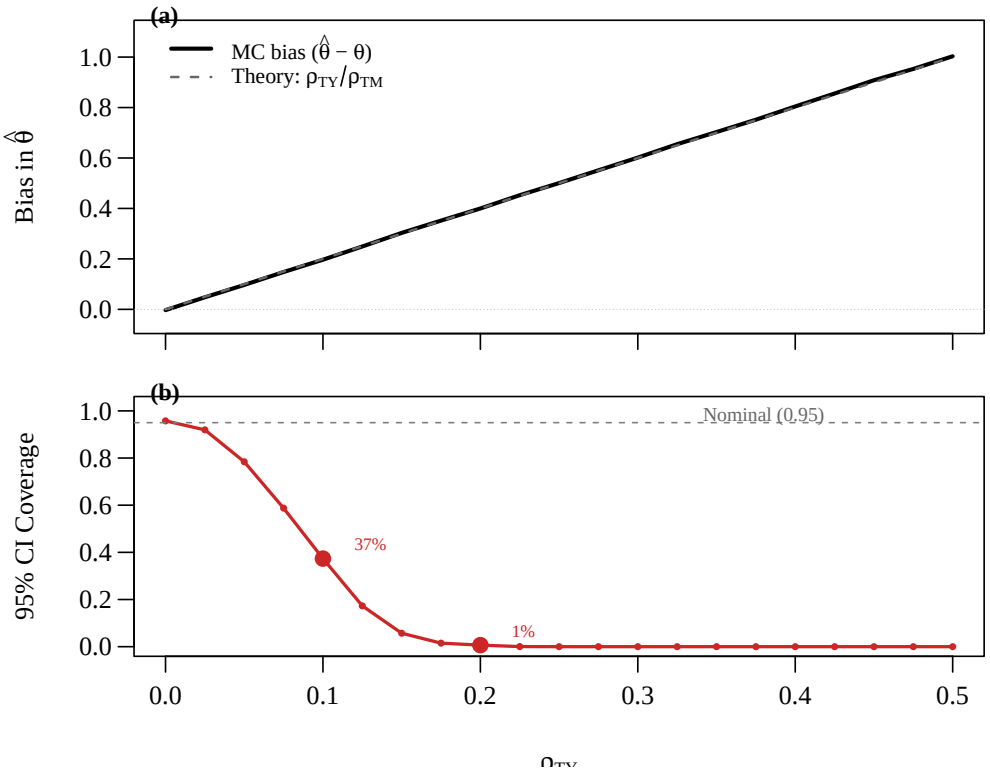


Figure 1: Bias and Coverage under MCA Violations

Notes: 2,000 replications at each grid point, $n = 2{,}500$, $\pi_1 = 0.5$, $\rho_{TM} = 0.5$. Panel (a): bias in $\hat{\theta}$ (solid black) with theoretical prediction ρ_{TY}/ρ_{TM} (dashed gray). Panel (b): empirical 95% CI coverage rate (solid red) with nominal level at 0.95 (dashed gray).

Size. Under MCA ($\rho_{TY} = 0$), all three tests reject at rates close to their nominal levels, confirming correct calibration.

Power. Table 6 reports rejection rates as a function of ρ_{TY}. Power increases monotonically in $|\rho_{TY}|$, reaching near-complete rejection by $\rho_{TY} = 0.20$. With one degree of overidentification, the J-statistic, squared t-statistic, and LR statistic are numerically identical, so we report a single rejection rate.[18]

$\hat{\kappa}$ Estimation. The LR test also delivers a point estimate $\hat{\kappa}$ of $\text{Cov}(\varepsilon_T, \varepsilon_Y)$ and a confidence interval. Under MCA, $\hat{\kappa}$ is centered at zero with nominal coverage. Under violations, $\hat{\kappa}$ tracks the true covariance with negligible bias and near-nominal coverage, providing a direct measure of departure from mediated confounding.

Table 6: Experiment 4: Size and Power of Specification Tests

	Rejection	$\hat{\kappa}$ Diagnostics		
ρ_{TY}	Rate at 5%	Bias($\hat{\kappa}$)	RMSE($\hat{\kappa}$)	Cov.($\hat{\kappa}$)
0	0.042	0.000	0.043	0.958
0.10	0.631	0.000	0.043	0.961
0.20	0.998	−0.002	0.043	0.958
0.30	1.000	−0.001	0.043	0.959
0.50	1.000	0.001	0.045	0.957

Notes: 2,000 replications with $n = 2{,}500$, $K = 2$ instruments ($\pi_1 = 0.5$, $\pi_2 = 0.3$, $\phi = 0.3$). Baseline: $\gamma = 0.4$, $\theta = 0.5$, $\tau = 0.3$, $\rho_{TM} = 0.5$, $\rho_{MY} = 0.3$. Overid dof $= 1$ (Section 5.1). With one degree of overidentification, the J-test, t^2-test, and LR test are numerically identical $\chi^2(1)$ statistics; a single rejection rate column is reported. Cov.($\hat{\kappa}$) = empirical 95% CI coverage rate for $\hat{\kappa}$.

7.6 Comparison with the Two-Instrument Benchmark

We compare the one-instrument (mediated confounding) approach with the two-instrument approach of Frölich and Huber (2017a). The two-instrument approach uses Z_1 as instrument for T and an independent Z_2 as instrument for M, estimating (τ, θ) by 2SLS of Y on (T, M) with instruments (Z_1, Z_2).

The DGP extends the baseline by adding a direct instrument for M: $M_i = \gamma T_i + \phi Z_{2i} + \varepsilon_{i,M}$, where $Z_{2i} \sim \mathcal{N}(0,1)$ is independent of Z_1 and of all errors. We set $\phi = 0.5$, $n = 2{,}500$, and $R = 2{,}000$.

Table 7 shows that both approaches are essentially unbiased with near-nominal coverage. The two-instrument estimator has smaller standard errors for $\hat{\theta}$, reflecting the stronger first stage for M provided by the direct instrument. The one-instrument estimator has slightly smaller standard errors for $\hat{\tau}$, as it exploits the additional moment condition $\text{E}[\varepsilon_T \varepsilon_Y] = 0$. The relative efficiency depends on the strength of the direct instrument (ϕ) and the collider mechanism (ρ_{TM}), and should be interpreted as illustrating the trade-off between assumption strength and statistical precision.

7.7 Robustness to Non-Gaussian Errors

The experiments above use Gaussian errors. Since identification relies on linearity and mean independence ($\text{E}[\varepsilon_T \varepsilon_Y] = 0$), not normality, we examine performance under heavy-tailed error distributions.

[18] This equivalence holds only when the overidentification degree of freedom equals one. With $K \geq 3$, the tests diverge in finite samples, as the t-test targets the MCA moment while the J-test pools across all restrictions.

Table 7: Comparison: One-Instrument vs. Two-Instrument Estimators

	Mediator Effect $\hat{\theta}$ (True = 0.5)				Direct Effect $\hat{\tau}$ (True = 0.3)			
Approach	Bias	RMSE	Cov.	SE	Bias	RMSE	Cov.	SE
One-IV (MCA)	−0.002	0.090	0.953	0.091	0.002	0.073	0.954	0.075
Two-IV (F&H)	−0.001	0.041	0.946	0.040	0.001	0.044	0.945	0.043

Notes: 2,000 replications, $n = 2{,}500$. One-IV uses Z_1 as instrument for M conditional on T (Section 4.2). Two-IV uses (Z_1, Z_2) as instruments for (T, M), with Z_2 entering the mediator equation directly ($\phi = 0.5$).

We replace the Gaussian error vector with a multivariate Student-t distribution with ν degrees of freedom, preserving the same correlation structure Σ as in the baseline DGP.[19] We consider $\nu = 5$ and $\nu = 3$, both under MCA ($\rho_{TY} = 0$) and under a moderate violation ($\rho_{TY} = 0.1$), using the baseline configuration ($n = 2{,}500$, $\pi_1 = 0.5$) with $R = 2{,}000$ replications.

Results. Table 8 reports the results alongside the Gaussian baseline. Under $t(5)$ errors, the estimator remains essentially unbiased, but coverage for $\hat{\theta}$ falls below nominal levels due to increased sampling variability relative to Gaussian-based standard errors. Under $t(3)$ errors, estimator performance deteriorates substantially, reflecting the sensitivity of 2SLS to very heavy tails. Under MCA violation ($\rho_{TY} = 0.1$), bias is similar across error distributions, confirming that the asymptotic bias formula $\hat{\theta} - \theta \approx \rho_{TY}/\rho_{TM}$ is distribution-free. These results suggest that, in heavy-tailed settings, robust or bootstrap inference should be used.

Table 8: Robustness to Non-Gaussian Errors

		Mediator Effect $\hat{\theta}$ (True = 0.5)				Direct Effect $\hat{\tau}$ (True = 0.3)			
Errors	ρ_{TY}	Bias	RMSE	Cov.	SE	Bias	RMSE	Cov.	SE
Panel A: MCA holds ($\rho_{TY} = 0$)									
Normal	0	−0.003	0.089	0.958	0.090	0.002	0.073	0.953	0.074
$t(5)$	0	−0.004	0.105	0.916	0.091	0.002	0.078	0.951	0.075
$t(3)$	0	0.098	4.335	0.712	4.013	−0.036	1.647	0.869	1.556
Panel B: MCA violated ($\rho_{TY} = 0.1$)									
Normal	0.1	0.197	0.215	0.374	0.087	−0.078	0.106	0.797	0.072
$t(5)$	0.1	0.202	0.225	0.356	0.087	−0.082	0.109	0.784	0.071

Notes: 2,000 replications, $n = 2{,}500$, $\pi_1 = 0.5$, $\rho_{TM} = 0.5$, $\rho_{MY} = 0.3$. Multivariate $t(\nu)$ errors preserve the same correlation structure as the Gaussian baseline; the scaling ensures unit marginal variance. Cov. = empirical 95% CI coverage rate. SE = average estimated standard error.

Summary. The comparison in Section 7.6 confirms that the one-instrument approach, under mediated confounding, achieves performance comparable to the two-instrument benchmark—without requiring a second instrument that directly shifts the mediator. The practical advantage is clear: the one-instrument design applies whenever a valid instrument for the treatment is available, which is vastly more common in empirical settings than having separate instruments for both the treatment and the mediator. The cost is the additional assumption of mediated confounding ($\rho_{TY} = 0$), which can be evaluated using

[19] Specifically, $(\varepsilon_T, \varepsilon_M, \varepsilon_Y)' = \mathbf{u}/\sqrt{V/\nu}$, where $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \Sigma \cdot (\nu - 2)/\nu)$ and $V \sim \chi^2(\nu)$ independently, yielding unit-variance $t(\nu)$ marginals.

three complementary diagnostics: the κ-based sensitivity analysis (Experiment 3), which quantifies how large a violation is needed to overturn the result; the R^2-based robustness values (Experiment 3), which bound reduced-form fragility to omitted confounders; and the specification tests (Experiment 4, requiring $K \geq 2$), which directly test $\kappa = 0$. The robustness check in Section 7.7 confirms that the method's validity extends beyond the Gaussian case to heavy-tailed error distributions, consistent with the theoretical requirement of mean independence rather than full independence.

8 Empirical Applications

In this section, we apply our IV mediation estimator to previous studies in the literature to showcase the utility of the estimator but to also highlight the relevance of the diagnostic statistics we provide. Table 9 reports the IV mediation estimates for five papers, as well as two applications that were used by Frölich and Huber (2017b). All papers and key variables are described in more detail in Appendix I.

Our first application is the study of Weber's Protestant work ethic hypothesis tested by Becker and Woessmann (2009). Using county-level data from Prussia on income taxes as a measure of economic prosperity, the share of Protestants as the treatment, and distance to Wittenberg as the instrument, they find a positive effect of Protestantism on taxes. Through a series of bounding exercises and 3SLS regressions, Becker and Woessmann (2009) argue that education mediates the entire effect of Protestantism on income taxes. Our IV mediation estimator reaches a similar conclusion. Our $\hat{\tau}^{\text{total}}$ replicates their main 2SLS result exactly (cf. Table V, column 4). The decomposition exhibits inconsistent mediation, with direct and indirect effects of opposite sign. The estimate $\hat{\tau}^{\text{direct}}$, which captures the residual effect of Protestantism on income taxes holding literacy fixed, is negative and statistically insignificant, while $\hat{\theta}\hat{\gamma}$, which captures the effect operating through literacy, is positive and larger than the total effect. The positive total effect is thus entirely accounted for by the education channel.[20]

Interestingly for us, several quantities in Table 9 compare to different exercises in Becker and Woessmann (2009). Their Table III, column 2 regresses the share literate on the share Protestant, instrumenting Protestantism with distance to Wittenberg. This is equivalent to estimating $\hat{\gamma}$, the effect of treatment on the mediator, for which they find a coefficient of 0.189 (s.e. = 0.028), compared to our estimate of 0.188 (s.e. = 0.027). In their 3SLS exercise in Table VIII, column 3, they find an effect of instrumented literacy on income taxes of 3.242 (s.e. = 1.169). This is analogous to estimating the effect of the mediator on the outcome, $\hat{\theta}$, and compares to our estimate of 4.709 (s.e. = 2.164).[21] Multiplying their estimates yields an implied indirect effect of 0.609, compared to our estimate of 0.886. This showcases how several tables in their paper are neatly summarized and rationalized by a single causal IV mediation estimator. In addition, their application generates strong instruments in both first stages and a very robust $\hat{\rho}^*_{TY}$ break-even point, which would require a sizable MCA violation.

The next two replications are Autor et al. (2020) and Dippel et al. (2022), which study the political consequences of trade-induced labor-market shocks in the United States and Germany. For Autor et al. (2020), our exercise is not an exact replication of their main paper because we augment their county-level political specification with a post-treatment mediator: the change in manufacturing employment

[20] The residual non-education channel, if anything, points in the opposite direction. This reinforces Becker and Woessmann's interpretation that Protestant economic advantage operated through human capital rather than a direct work-ethic effect.

[21] Unlike our estimator, their 3SLS system does not estimate the effect of the mediator on the outcome while directly controlling for Protestantism.

share from 2000 to 2007, taken from the original China-shock study in Autor et al. (2013). For Dippel et al. (2022), we use their Kreis-level stacked panel and the change in employment in trade-dependent manufacturing sectors as the mediator.[22] In both settings, the IV mediation estimator delivers a qualitatively similar conclusion: import competition increases support for right-wing parties, and this effect is largely or entirely mediated by labor-market disruption. Despite differences in country context, outcome measurement, treatment scaling, and geographic units, the decomposition is remarkably consistent across the two applications. In the U.S. case, the labor-market channel more than accounts for the total effect, implying a negative residual direct effect. In the German case, the indirect effect is approximately equal to the total effect and the direct effect is close to zero. Thus, across two distinct trade-and-politics settings, the IV mediation estimator summarizes the evidence in a common way: the political response to import competition operates primarily through adverse local labor-market consequences rather than through residual direct channels.

Application four is based on work by Devoto et al. (2012). They study demand for private household water connections in Morocco when connections can be financed through credit. Here, Y is an indicator for whether the household reports that life improved after 24 months, T is take-up of a private water connection, Z is assignment to treatment, and M is a leisure and social well-being index. This application provides a useful contrast to the previous cases: while the total IV effect is positive and precisely estimated,[23] $\hat{\tau}^{\text{total}} = 0.348$ (s.e. = 0.057), the estimated mediator effect is close to zero, $\hat{\theta} = -0.034$ (s.e. = 0.691), yielding a negligible indirect effect of -0.002 (s.e. = 0.049). Instead, the direct effect, $\hat{\tau}^{\text{direct}} = 0.351$ (s.e. = 0.036), is nearly identical to the total effect. The conditional mediator first stage is relatively weak, $F_{MZ|T} = 5.21$, so the absence of a mediated effect should be interpreted cautiously. Nevertheless, the decomposition provides little evidence that the measured leisure and social well-being index mediates the effect of private water connections on reported life improvement; the estimated effect is driven almost entirely by the direct channel.

The next application is Arnsbarger et al. (2026), who study whether women's labor-market participation enabled political mobilization after the U.S. Civil War. Here, Y is an indicator for Temperance Crusade protest activity in 1873, T is the standardized share of disabled or wounded Union Army veterans, Z is the share of soldiers excluded from combat, and M is standardized female labor-force participation in 1870. The IV mediation estimator delivers a conceptually natural result: the total effect is positive, $\hat{\tau}^{\text{total}} = 0.038$ (s.e. = 0.010),[24] the direct effect is essentially zero, $\hat{\tau}^{\text{direct}} = -0.002$ (s.e. = 0.005), and the indirect effect through female labor-force participation accounts for the full effect, $\hat{\theta}\hat{\gamma} = 0.040$ (s.e. = 0.011). Both first stages are reasonably strong, with $F_{TZ} = 422.02$ and $F_{MZ|T} = 12.65$. The result is somewhat less robust than Becker and Woessmann according to $\hat{\rho}^{*}_{TY}$, but still provides clear evidence of a plausible mediated channel.

The final two applications revisit the empirical illustrations in Frölich and Huber (2017b). Their paper studies mediation using two distinct instruments, one for the treatment and one for the mediator, whereas our estimator uses a single instrument. Despite this difference, we recover qualitatively similar conclusions. In the BHPS application, where the treatment is education, the outcome is social functioning, and the mediator is income, Frölich and Huber (2017b) estimate a LATE of 3.272, a parametric direct effect of 3.397, and a parametric indirect effect of -0.029. Our estimates are similar in

[22] Our $\hat{\tau}^{\text{total}}$ estimate replicates the result in their Table 1, column 5.
[23] Our $\hat{\tau}^{\text{total}}$ estimate replicates the result in their Table 11, panel b, column 4.
[24] Our $\hat{\tau}^{\text{total}}$ estimate replicates the result in their Table 6, column 4.

interpretation: $\hat{\tau}^{\text{total}} = 3.274$ (s.e. = 1.037), with most of the effect operating through the direct channel, $\hat{\tau}^{\text{direct}} = 2.609$ (s.e. = 2.024), rather than through income, $\hat{\theta}\hat{\gamma} = 0.665$ (s.e. = 1.953). In the Job Corps application, Frölich and Huber (2017b) estimate a LATE of 12.797, a parametric direct effect of -0.824, and a parametric indirect effect of 13.188. We find the same qualitative pattern: the direct effect is close to zero, $\hat{\tau}^{\text{direct}} = -0.410$ (s.e. = 18.243), while the indirect effect, $\hat{\theta}\hat{\gamma} = 12.956$ (s.e. = 21.567), is nearly identical to the total effect of 12.546. The main caveat is that neither application has a strong conditional first stage for the mediator, with $F_{MZ|T} = 1.43$ in BHPS and $F_{MZ|T} = 0.15$ in Job Corps. The sensitivity diagnostics also suggest that the BHPS result is more robust than the Job Corps result, with $\hat{\rho}^*_{TY} = -0.220$ compared to -0.015. Thus, although our estimator uses only one instrument, it reproduces the qualitative conclusions of their two-instrument mediation applications: income does not mediate the education effect in BHPS, whereas hours worked account for the Job Corps earnings effect.

Table 9: IV Mediation Applications to Other Papers

	Becker & Woessmann (2009)	Autor et al. (2020)	Dippel et al. (2022)	Devoto et al. (2012)	Arnsbarger et al. (2026)	Froelich & Huber (2017) BHPS	Froelich & Huber (2017) Job Corps
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Mediation estimates							
$\hat{\tau}^{\text{total}}$ (Wald)	0.586	1.588	0.092	0.348	0.038	3.274	12.546
	(0.242)	(0.857)	(0.044)	(0.057)	(0.010)	(1.037)	(6.116)
$\hat{\tau}$ (direct $T \to Y$)	-0.299	-0.955	-0.002	0.351	-0.002	2.609	-0.410
	(0.214)	(0.547)	(0.033)	(0.036)	(0.005)	(2.024)	(18.243)
$\hat{\theta}$ (mediator $\to Y$)	4.709	-1.024	-3.927	-0.034	0.304	-0.000	6.676
	(2.164)	(0.443)	(1.845)	(0.691)	(0.117)	(0.000)	(10.631)
$\hat{\gamma}$ ($T \to$ mediator)	0.188	-2.483	-0.024	0.071	0.131	-1654.289	1.941
	(0.027)	(0.398)	(0.009)	(0.034)	(0.031)	(6191.048)	(0.744)
$\hat{\theta}\hat{\gamma}$ (indirect)	0.886	2.542	0.094	-0.002	0.040	0.665	12.956
	(0.408)	(1.110)	(0.059)	(0.049)	(0.011)	(1.953)	(21.567)
Panel B: Diagnostics							
F_{TZ} (standard)	75.69	82.13	64.33	407.44	422.02	26.27	6,046
$F_{MZ\|T}$ (conditional)	16.36	31.20	8.93	5.21	12.65	1.43	0.15
Panel C: Sensitivity							
$\hat{\rho}^*_{TY}$ (break-even)	-0.257	-0.169	-0.095	0.003	-0.096	-0.220	-0.015
n	426	3,107	730	788	8,882	3,428	4,603
Clusters	—	722	93	592	202	—	—

9 Conclusion

When a valid instrument identifies the total effect of a treatment T on an outcome Y, can the same instrument decompose that effect into a direct component and an indirect component operating through a mediator M? The answer is negative under standard IV assumptions (Section 2), but affirmative under the mediated confounding assumption—the requirement that the unobserved determinants of treatment

be independent of the unobserved determinants of the outcome ($\varepsilon_T \perp\!\!\!\perp \varepsilon_Y$), while allowing the mediator error to correlate freely with both (MCA (14)).

The mediated confounding assumption generates two key properties: conditional exogeneity and conditional relevance of the instrument for the mediator given treatment (Proposition 1 and Section 3). In the linear model, these properties deliver closed-form identification of all mediation parameters via four moment conditions (Theorem 1), and the mediation decomposition is internally consistent with the standard Wald ratio (Corollary 1). The analysis extends to nonlinear settings through the local instrumental variable approach, where mediation effects are identified by the second derivative of conditional expectations with respect to the propensity score (Theorem 2), and to the binary instrument case through a conditional Wald ratio that identifies a local mediator effect within each treatment stratum under MCA, treatment monotonicity, and a compositional mediator monotonicity condition weaker than the standard requirement $M(1) \geq M(0)$ (Theorem 3).

The method has several practical advantages. It requires only one instrument—the same one already used for the treatment effect. It permits both T and M to be endogenous, avoiding sequential ignorability. It ensures that direct and indirect effects correspond to the same complier population. And it can be implemented using standard 2SLS software. We complement the estimator with specification tests for mediated confounding under overidentification and a sensitivity analysis—based on robustness values adapted from the omitted variable bias framework—that applies even with a single instrument.

The mediated confounding assumption is restrictive, and the diagnostic and sensitivity tools developed in Section 5 are essential for assessing its plausibility in any given application. A further practical concern is that the conditional first stage is powered by the collider mechanism rather than by a direct instrument-mediator channel, making the method more susceptible to weak instrument problems than standard IV. Applied researchers should report the conditional first-stage F-statistic alongside the robustness values to assess whether sufficient variation is available for reliable inference.

Several extensions merit further investigation. The approach currently handles a single mediator; extending it to multiple, possibly sequential mediators would broaden applicability to settings with complex causal chains. Developing nonparametric or semiparametric estimators for the second-derivative identification result—and characterizing their finite-sample properties—is an important next step. Appendix H shows that the linear identification result extends to settings with high-dimensional candidate controls through a quadruple-LASSO post-double selection procedure: four parallel LASSO regressions select the relevant components of a rich expansion of the controls for Y, T, M, and the instrument Z, and the resulting post-selection 2SLS preserves $\sqrt{n}$-consistency, asymptotic normality, the conditional first-stage diagnostics of Remark 2, and the κ-based sensitivity analysis of Section 5. The framework is fully compatible with cross-fitting in the sense of Chernozhukov et al. (2018).

References

Albert, Jeffrey M., "Mediation analysis via potential outcomes models," *Statistics in Medicine*, 2008, *27* (8), 1282–1304.

Anderson, T. W. and Herman Rubin, "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 1949, *20* (1), 46–63.

Arnsbarger, Madison, Andreas Ferrara, and Paige Montrose, "The U.S. Civil War's Impact on Women's Work and Political Participation," *working paper*, 2026.

Attanasio, Orazio, Sarah Cattan, Emla Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina, "Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia.," *American Economic Review*, 2020, *110*, 48–85.

Autor, David, David Dorn, and Gordon Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review*, 2013, *103* (6), 2121–2168.

_ , _ , _ , and Kaveh Majlesi, "Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure," *American Economic Review*, 2020, *110* (10), 3139–3183.

Becker, Sascha O. and Ludger Woessmann, "Was Weber Wrong? A Human Captial Theory of Protestant Economic History," *Quarterly Journal of Economics*, 2009, *124* (2).

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen, "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 2012, *80* (6), 2369–2429.

_ , Victor Chernozhukov, and Christian Hansen, "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 2014, *81* (2), 608–650.

_ , _ , _ , and Kengo Kato, "High-Dimensional Econometrics and Regularized GMM," arXiv:1806.01888 2018. Available at https://arxiv.org/abs/1806.01888.

Brunello, Giorgio, Margherita Fort, Nicole Schneeweis, and Rudolf Winter-Ebmer, "The Causal Effect of Education on Health: What is the Role of Health Behaviors?," *Health Economics*, 2016, *25*, 314–336.

Chen, Stacey H., Yen-Chien Chen, and Jin-Tan Liu, "The Impact of Family Composition on Educational Achievement," *Journal of Human Resources*, 2019, *54* (1), 122–170.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 2018, *21* (1), C1–C68.

Cinelli, Carlos and Chad Hazlett, "Making Sense of Sensitivity: Extending Omitted Variable Bias," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2020, *82* (1), 39–67.

_ and _ , "An Omitted Variable Bias Framework for Sensitivity Analysis of Instrumental Variables," *Quantitative Economics*, 2024, *15* (3), 723–765. Working paper version: 2022.

Dawid, A. P., "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society, Series B*, 1979, *41* (1), 1–31.

Devoto, Florencia, Esther Duflo, Pasqualine Dupas, William Parienté, and Vincent Pons, "Happiness on Tap: Piped Water Adoption in Urban Morocco," *American Economic Journal: Economic Policy*, 2012, *4* (4), 68–99.

Dieterle, Steven G. and Andy Snell, "A Simple Diagnostic to Investigate Instrument Validity and Heterogeneous Effects When Using a Single Instrument," *Labour Economics*, 2016, *42*, 76–86.

Dippel, Christian, Robert Gold, and Stephan Heblich, "The Effect of Trade on Workers and Voters," *Economic Journal*, 2022, *132* (641), 199–217.

Dunn, Graham and Richard Bentall, "Modelling treatment-effect heterogeneity in randomized con-

trolled trials of complex interventions (psychological treatments)," *Statistics in Medicine*, 2007, *26* (26), 4719–4745.

Fan, Jianqing and Irène Gijbels, *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall/CRC, 1996.

Frölich, Markus and Martin Huber, "Direct and Indirect Treatment Effects–Causal Chains and Mediation Analysis with Instrumental Variables," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2017, *79* (5), 1645–1666.

Frölich, Markus and Martin Huber, "Direct and Indirect Treatment Effects–Causal Chains and Mediation Analysis with Instrumental Variables," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2017, *79* (5), 1645–1666.

Hausman, J. A., "Specification Tests in Econometrics," *Econometrica*, 1978, *46* (6), 1251–1271.

Hayashi, Fumio, *Econometrics*, Princeton, NJ: Princeton University Press, 2000.

Heckman, James J. and Edward J. Vytlacil, "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, April 1999, *96* (8), 4730–4734.

_ and _ , "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *73* (3), 669–738.

_ , Rodrigo Pinto, and Peter A. Savelyev, "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, October 2013, *103* (6), 2052–2086.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto, "Experimental Designs for Identifying Causal Mechanisms," *Journal of the Royal Statistical Society A*, 2013, *176* (1), 5–51.

_ , Luke Keele, and Teppei Yamamoto, "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 2010, *25* (1), 51–71.

Imbens, Guido W. and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, March 1994, *62* (2), 467–475.

Joffe, Marshall M., Dylan Small, Thomas T. Have, Steve Brunelli, and Harold I. Feldman, "Extended Instrumental Variables Estimation for Overall Effects," *International Journal of Biostatistics*, 2008, *4* (1).

Kitagawa, Toru, "A Test for Instrument Validity," *Econometrica*, 2015, *83* (5), 2043–2063.

Kwon, Soonwoo and Jonathan Roth, "Testing Mechanisms," *Review of Economic Studies*, 2026. Conditionally accepted.

Mattei, Alessandra and Fabrizia Mealli, "Augmented designs to assess principal strata direct effects," *Journal of the Royal Statistical Society B*, 2011, *73* (5), 729–752.

Navjeevan, Manu, Rodrigo Pinto, and Andres Santos, "Identification in Instrumental Variables Models: The Central Role of Abadie's Kappa," *Econometrica*, 2026. Forthcoming.

Olea, José Luis Montiel and Carolin Pflueger, "A Robust Test for Weak Instruments," *Journal of Business & Economic Statistics*, 2013, *31* (3), 358–369.

Robins, James M. and Sander Greenland, "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 1992, *3* (2), 143–155.

Rudolph, Kara E., Nicholas Williams, and Iván Díaz, "Using instrumental variables to address unmeasured confounding in causal mediation analysis," *Biometrics*, 2024, *80* (1), 1–11.

Small, Dylan S., "Mediation analysis without sequential ignorability: using baseline covariates interacted with random assignment as instrumental variables," *Journal of Statistical Research*, 2012, *46* (2), 91–103.