Mediation Analysis With a Single Instrument

Preliminary and Incomplete, do not cite *

Christian Dippel† Robert Gold‡ Stephan Heblich§ Rodrigo Pinto¶

April 12, 2019

Abstract

The method of Two-stage Leas Squares (2SLS) employs instrumental variables (IV) to evaluate the causal effects of an endogenous treatment on an outcome of interest. The method is widely adopted by empirical economists who typically investigate a data set consisting of the instrument variable, the treatment variable and multiple outcomes. Although a single instrument enables the identification of treatment effects on multiple outcomes, it cannot identify the causal effect of one outcome on another. In particular, the standard IV model cannot unpack the causal effects that arise when the treatment variable and an outcome together cause a second outcome of interest. We present a simple identification strategy that enables the researcher to use instrumental variables to identify the causal relation among outcomes. Our method does not require additional instruments. It exploits an identifying assumption regarding unobserved error terms that maintains the endogeneity of the treatment variable while allowing for endogeneity among outcomes. This paper offers a novel application of the well-known method of 2SLS for a class of IV model with a single instrument and multiple outcomes.

Keywords: Instrumental Variables, Causal Mediation Analysis,
JEL Codes: C36

*We thank Sonia Bhalotra, Johanna Fajardo, Andreas Ferrara, Markus Frölich, James Heckman, Martin Huber, Kosuke Imai, Ed Leamer, Yi Lu, Craig McIntosh, Bruno Pellegrino, David Slichter, Dustin Tingley, Frank Windmeijer for valuable discussions. We also thank David Slichter for thoughtful comments.

†University of California, Los Angeles, CCPR, and NBER.

‡IfW - Kiel Institute for the World Economy and CESifo.

§University of Bristol, CESifo, IZA, and SERC.

¶University of California, Los Angeles, CCPR, and NBER.

1 Introduction

A basic inquiry in policy evaluation is to identify the causal effect of a treatment variable on an outcome of interest. A primary problem to assess causal effects using observational data is endogeneity, when the treatment and the outcome are confounded by unobserved variables. Endogeneity induces a correlation between the treatment and the outcome which hinders the identification of treatment effects.[1] Economists have long used instrumental variables (IV) to solve this identification problem. The stan that inflicts endogenous treatments.

$$T = f_T(Z, \boldsymbol{V}, \epsilon_T) \tag{1}$$

$$Y = f_Y(T, \boldsymbol{V}, \epsilon_Y) \tag{2}$$

This

Standard IV estimation, however, is unable to unpack the causal chain that arises when the treatment and its outcome jointly cause a second outcome of interest.

We investigate the problem of identifying causal relations when an endogenous treatment and its outcome together cause a second outcome of interest. We propose a solution to the problem that does not require additional instrumental variables and can be easily implemented using the well-known two-stage least squares (2SLS) estimator.

The starting point is to estimate the effect of a non-random treatment T (e.g. TBA) on an outcome M (e.g. TBA). The solution involves using an instrumental variable Z that affects T (i.e. there is a first-stage relation) but is uncorrelated with the omitted variables (i.e. the exclusion restriction holds).

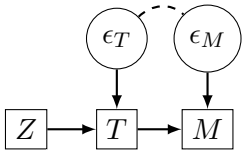
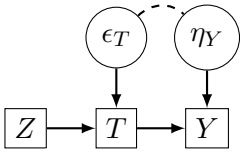
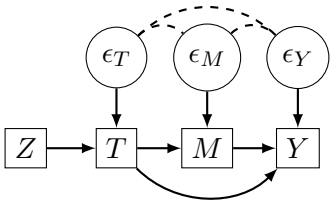
This is the standard IV solution and is depicted in *Model I* in Table 1. T is endogenous in a regression of M on T (i.e. $\epsilon_T \not\perp\!\!\!\perp \epsilon_M$), but Z is exogenous (i.e. $Z \perp\!\!\!\perp \epsilon_T, \epsilon_M$).

We are interested in the identification challenge that arises when there is a second outcome of interest Y (e.g. TBA) that is caused by T, both through M as well as "directly" (which is to say, through other channels). The most straightforward approach to this is to simply estimate the 'total

[1] Common sources of endoegenity are selection bias, omitted variables, measurement error, sample selection, of functional form misspecifications.

Table 1: The Identification Problem of Mediation Analysis with IV

A. Graphical Representation

Model I: *IV for M*	*Model II:* *IV for Y*	*Model III:* *IV for the Mediation Model*
		

B. Model Equations

$T = f_T(Z, \epsilon_T)$	$T = f_T(Z, \epsilon_T)$	$T = f_T(Z, \epsilon_T),\ M = f_M(T, \epsilon_M)$
$M = f_M(T, \epsilon_M)$	$Y = g_Y(T, \eta_Y)$	$Y = f_Y(T, M, \epsilon_Y)$
$Z \perp\!\!\!\perp (\epsilon_T, \epsilon_M)$	$Z \perp\!\!\!\perp (\epsilon_T, \eta_Y)$	$Z \perp\!\!\!\perp (\epsilon_T, \epsilon_M, \epsilon_Y)$

Notes: (*a*) *Model I* is the standard IV model, which enables the identification of the causal effect of T on M. *Model II* is the standard IV model that enables the identification of the causal effects of T on Y. *Model III* is the IV Mediation Model with an instrumental variable Z. (*b*) Panel A gives the graphical representation of the models. Panel B presents the non-parametric structural equations of each model. Conditioning variables are suppressed for sake of notational simplicity. We use $\perp\!\!\!\perp$ to denote statistical independence.

effect' of T on Y using the same IV approach, as depicted in *Model II* in Table 1:

$\epsilon_T \not\perp\!\!\!\perp \eta_Y$, but Z is exogenous (i.e. $Z \perp\!\!\!\perp \epsilon_T, \eta_Y$).[2]

In combination, *Model I* and *Model II* estimate the causal effect of T on M and the causal effect of T on Y. However, this does not identify to what extent the former causes the latter.

The identification challenges that arise from this discussion are depicted in *Model III* in Table 1. Equations $M = f_M(T, \epsilon_M)$ and $Y = f_Y(T, M, \epsilon_Y)$ imply that T causes Y indirectly through M as well as directly, i.e. through other channels (that are graphically represented by the arrow directly linking T to Y).

In a regression of Y on both T and M, there are two endogenous regressors (i.e. $\epsilon_T \not\perp\!\!\!\perp \epsilon_Y$, $\epsilon_M \not\perp\!\!\!\perp \epsilon_Y$), but there is only one instrument Z to address this endogeneity. *Model III* is a *mediation model*, i.e. one where T causes an intermediate outcome M that is also a *mediator* in T's effect on a final outcome Y.

[2]It is common to use the same instrument to identify the causal effect of a treatment on several outcomes, and the application studied here is no different. For example, in the literature investigating the effect of trade shocks on local labor markets, three pairs of well-known and cited papers each use the IV strategy to separately investigate the effect of trade on labor markets and on some form of political outcomes; e.g. Autor et al. (2013) and Autor et al. (2016), Malgouyres (2017) and Malgouyres (2014), as well as Pierce and Schott (2016) and Che et al. (2016).

[3]

Most of the approaches to identification in mediation analysis assume that T is as good as randomly assigned (i.e. $\epsilon_T \perp\!\!\!\perp \epsilon_M$), making them not applicable to the IV settings we are interested in. See, e.g., Imai et al. (2011). The only existing approaches to achieving identification in the IV setting of *Model III* require separate dedicated instruments for M, which require additional exogeneity assumptions that are considerably more restrictive than the standard ones (e.g. Frölich and Huber 2017; Jun et al. 2016).

Our proposed solution does not assume away endogeneity in any of the key relationships in *Model III* and does not require additional instruments. Instead, we rely on the insight that in many research settings the omitted variable concerns themselves suggest a natural solution.

This is the case when T is endogenous in a regression of Y on T primarily because of omitted variables that affect M. We show that this assumption alone is sufficient to unpack the causal channels in *Model III*, allowing us to identify the extent to which T causes Y through M.

We further show that under linearity, the resulting identification framework is straightforwardly estimated using three separate 2SLS estimations of the effect of T on M, the effect of T on Y, and the effect of M on Y conditional on T.

We also develop a procedure to bound the possible range of the direct and the indirect effects linking T and Y when the identifying assumption of our framework is relaxed.

Our paper makes a methodological contribution to the literature on causal mechanisms and on IV. We offer a mediation model that relies on a single instrumental variable Z that directly causes T to identify three causal effects, while allowing for endogenous variables caused by confounders and for unobserved mediators. This parsimonious feature is useful for the typical observational data setting, where good instrumental variables are scarce.

Our model can be estimated by well-known 2SLS methods, its identifying assumption can be relaxed to derive bounds instead of point estimates, and it can be applied to a potentially broad range of empirical research questions in which an endogenous treatment and its primary outcome together cause a second outcome of interest.

The rest of the paper proceeds as follows: Section **??** explains our identification approach. We

[3] Mediation analysis decomposes the *total effect* of T on Y into the *indirect effect* of T on Y that operates through M and the *direct effect* that does not. The indirect effect may alternatively be labeled as the '*mediated effect*'. For recent works on this literature, see Heckman and Pinto (2015); Imai et al. (2010); Pearl (2014).

refer the reader to section **??** for an informal discussion of the model's intuition. Section **??** presents the IV results for *Model I* and *Model II*, establishing the causal effects of import exposure on labor markets and voting. Section **??** describe a simulated model. Section **??** concludes.

References

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), 2121–68.

_ , _ , _ , and Kaveh Majlesi, "Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure," *NBER Working Paper*, 2016.

Che, Yi, Yi Lu, Justin R Pierce, Peter K Schott, and Zhigang Tao, "Does Trade Liberalization with China Influence US Elections?," Technical Report, National Bureau of Economic Research 2016.

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, pp. n/a–n/a.

Heckman, James J and Rodrigo Pinto, "Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs," *Econometric reviews*, 2015, *34* (1-2), 6–31.

Imai, Kosuke, Luke Keele, and Dustin Tingley, "A General Approach to Causal Mediation Analysis," *Psychological Methods*, 2010, *15* (4), 309–334.

_ , _ , _ , and Teppei Yamamoto, "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies," *American Political Science Review*, 2011, *105* (4), 765–789.

Jun, Sung Jae, Joris Pinkse, Haiqing Xu, and Neşe Yildiz, "Multiple Discrete Endogenous Variables in Weakly-Separable Triangular Models," *Econometrics*, 2016, *4* (1).

Malgouyres, Clement, "The Impact of Exposure to Low-Wage Country Competition on Votes for the Far-Right: Evidence from French Presidential Elections," *working paper*, 2014.

Malgouyres, Clément, "The impact of chinese import competition on the local structure of employment and wages: Evidence from france," *Journal of Regional Science*, 2017, *57* (3), 411–441.

Pearl, Judea, "Interpretation and identification of causal mediation," *Psychological Methods, Special Section: Naturally Occurring Section on Causation Topics in Psychological Methods*, 2014, *19*, 459–481.

Pierce, Justin R and Peter K Schott, "The Surprisingly Swift Decline of US Manufacturing Employment," *American Economic Review*, 2016, *106* (7), 1632–62.

Appendix A Proofs of Exclusion Restrictions

Proof. The treatment equation $T = f_T(Z, \epsilon_T)$ in equation (**??**) implies that $Z \not\perp\!\!\!\perp T$. Thus our task is to prove two exclusion restrictions: $Z \perp\!\!\!\perp M(t)$ and $Z \perp\!\!\!\perp Y(t)$. According to (**??**), the counterfactual mediation is given by $M(t) = f_M(t, \epsilon_M)$. But Assumption **??** states that $Z \perp\!\!\!\perp (\epsilon_T, \epsilon_M, \epsilon_Y)$. In particular, we have that:

$$Z \perp\!\!\!\perp \epsilon_T \Rightarrow Z \perp\!\!\!\perp f_M(t, \epsilon_M) \Rightarrow Z \perp\!\!\!\perp M(t). \tag{3}$$

We can use iterated substitution to express the outcome counterfactual $Y(t)$ in equation (**??**) as the following function of error terms:

$$Y(t) = f_Y(t, M(t), \epsilon_Y) = f_Y(t, f_M(t, \epsilon_M), \epsilon_Y) \text{ by } (\textbf{??}), \tag{4}$$

$$\text{by } \textbf{??} \text{ we have that: } Z \perp\!\!\!\perp (\epsilon_M, \epsilon_Y) \Rightarrow Z \perp\!\!\!\perp f_Y(t, f_M(t, \epsilon_M), \epsilon_Y) \Rightarrow Z \perp\!\!\!\perp Y(t). \tag{5}$$

□

Proof. The lemma requires two proofs. The first shows that Z is not independent of M conditioned on T, that is, $Z \not\perp\!\!\!\perp M|T$. The second shows that the exclusion restriction $Z \perp\!\!\!\perp Y(m)|T$ holds under the independence condition $\epsilon_T \perp\!\!\!\perp \epsilon_Y$ in Assumption **??**.

An intuitive justification for $Z \not\perp\!\!\!\perp M|T$ relies on interpreting the correlations generated by condition on T. Recall that the treatment equation is given by $T = f_T(Z, \epsilon_T)$. Thus, conditioning on $T = t$ is equivalent to conditioning on the values of Z, ϵ_T such that $f_T(Z, \epsilon_T) = t$. This induces a correlation between Z and ϵ_T and thereby $Z \not\perp\!\!\!\perp \epsilon_T|T$. Moreover, ϵ_T correlates with ϵ_M and therefore we also have that $Z \not\perp\!\!\!\perp \epsilon_M|T$. But if $Z \not\perp\!\!\!\perp \epsilon_M|T$, then $Z \not\perp\!\!\!\perp f_M(T, \epsilon_M)|T$ and therefore we have that $Z \not\perp\!\!\!\perp M|T$ as $M = f_M(T, \epsilon_M)$. In summary, conditioning on T induces a correlation between Z and ϵ_T, but error term ϵ_T correlates with ϵ_M, which in turn generates a correlation between Z and M. It remains to be shown that the independence relation $\epsilon_T \perp\!\!\!\perp \epsilon_Y$ generates the exclusion restriction $Z \perp\!\!\!\perp Y(m)|T$, where the outcome counterfactual $Y(m)$ is given by $Y(m) = f_Y(T, m, \epsilon_Y)$ as in equation (**??**). The following rationale justify this assessment. Assumptions **??**–**??** generates the unconditional independence relation $(\epsilon_T, \epsilon_Z) \perp\!\!\!\perp \epsilon_Y$. Let $f_1(\cdot), f_2(\cdot), f_3(\cdot)$ be three arbitrary non-degenerate functions such that $f_1 : \text{supp}(\epsilon_Z) \times \text{supp}(\epsilon_T) \to \mathbb{R}$, $f_2 : \text{supp}(\epsilon_Z) \to \mathbb{R}$,

$f_3 : \text{supp}(\epsilon_Y) \to \mathbb{R}$. Under this notation, we have that:

$$(\epsilon_T, \epsilon_Z) \perp\!\!\!\perp \epsilon_Y \quad \Rightarrow \quad \epsilon_Z \perp\!\!\!\perp \epsilon_Y | f_1(\epsilon_Z, \epsilon_T) \quad \Rightarrow \quad f_2(\epsilon_Z) \perp\!\!\!\perp f_3(\epsilon_Y) | f_1(\epsilon_Z, \epsilon_T). \tag{6}$$

In particular, we can set functions $f_1(\epsilon_T), f_2(\epsilon_Y), f_3(\epsilon_Z, \epsilon_T)$ in (6) to the following expressions: $f_1(\epsilon_Z) = f_Z(\epsilon_Z)$, $f_2(\epsilon_Y) = f_Y(t, m, \epsilon_Y)$, and $f_3(\epsilon_Z, \epsilon_T) = f_T(f_Z(\epsilon_T), \epsilon_Z)$. Thus:

$$f_2(\epsilon_Z) \perp\!\!\!\perp f_3(\epsilon_Y) \mid f_1(\epsilon_Z, \epsilon_T) \tag{7}$$

$$\Rightarrow f_Z(\epsilon_Z) \perp\!\!\!\perp f_Y(t, m, \epsilon_Y) \mid \big(f_T(f_Z(\epsilon_T), \epsilon_Z) = t\big) \quad \forall (t, m) \in \text{supp}(T) \times \text{supp}(M) \tag{8}$$

$$\Rightarrow Z \perp\!\!\!\perp f_Y(t, m, \epsilon_Y) \mid \big(T = t\big) \quad \forall (t, m) \in \text{supp}(T) \times \text{supp}(M) \tag{9}$$

$$\Rightarrow Z \perp\!\!\!\perp Y(m) | T. \tag{10}$$

□

Proof. Assumptions **??**–**??** implies that $\epsilon_Y \perp\!\!\!\perp (Z, \epsilon_T)$. According to equation (**??**), we have that:

$$\begin{aligned} P(Y(m) \leq y | T = t) &= P(f_Y(t, m, \epsilon_Y) \leq y | T = t), \\ &= P(f_Y(t, m, \epsilon_Y) \leq y | f_T(Z, \epsilon_T) = t), \\ &= P(f_Y(t, m, \epsilon_Y) \leq y), \\ &= P(Y(t, m) \leq y), \end{aligned}$$

where the third equality comes from $\epsilon_Y \perp\!\!\!\perp (Z, \epsilon_T)$. □

Appendix B Generating the Error Structure and in Simulated Data

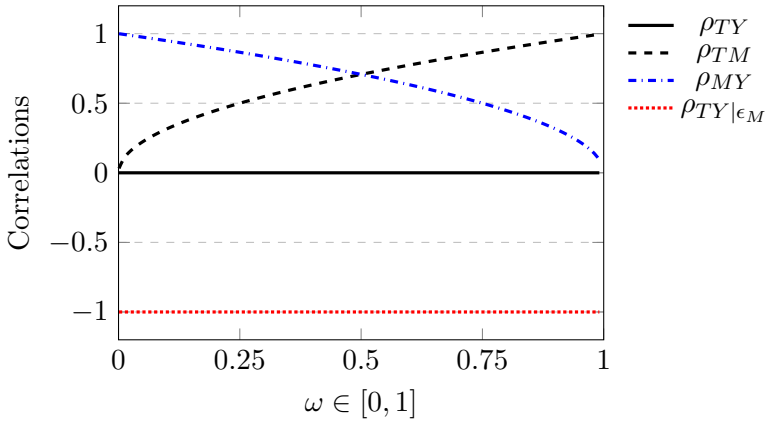
It is instructive to show how the dependence relations in Assumption **??** translate into ρ_{TM}, ρ_{MY}, and ρ_{TY} in Σ_ϵ, equation (**??**). A simulated dataset with the dependence relations in **??** can be straightforwardly generated in the following way:

- Separately generate error terms ϵ_T, ϵ_Y that are normally distributed with mean zero and variance one, $N(0,1)$. These are statistically independent, i.e. $\epsilon_T \perp\!\!\!\perp \epsilon_Y$.
- Let error term ϵ_M be defined as $\epsilon_M = \sqrt{\omega} \cdot \epsilon_T + \sqrt{(1-\omega)} \cdot \epsilon_Y$ for any $\omega \in [0,1]$.[4]

The correlation between ϵ_M, ϵ_T is given by $\rho_{TM} = \sqrt{\omega}$. Thereby $\epsilon_M \not\perp\!\!\!\perp \epsilon_T$. By symmetry, we also have that $\rho_{MY} = \sqrt{(1-\omega)}$ and $\epsilon_M \not\perp\!\!\!\perp \epsilon_Y$. Having drawn ϵ_T, ϵ_Y independently implies that the correlation between ϵ_T and ϵ_Y is $\rho_{TY} = 0$. However, conditioning on $\epsilon_M = e$ induces a linear relation between ϵ_T, ϵ_Y, namely, $\epsilon_T = e/\sqrt{\omega} - \sqrt{(1-\omega)/\omega} \cdot \epsilon_Y$. Thus, the correlation between ϵ_T, ϵ_Y conditioned on ϵ_M is $\rho_{TY|\epsilon_M} = -1$ and thereby $\epsilon_T \not\perp\!\!\!\perp \epsilon_Y | \epsilon_M$. Figure A1 displays the model correlations $\rho_{TM}, \rho_{MY}, \rho_{TY}$ and $\rho_{TY|\epsilon_M}$ as a function of $\omega \in [0,1]$. A high ω implies a high ρ_{TM}. By contrast, a low ω implies a high ρ_{MY}.

[4] Note that $\epsilon_T \sim N(0,1)$ and $\epsilon_Y \sim N(0,1)$ imply $\epsilon_M \sim N(0,1)$.

Figure A1: Correlations Among Error Terms by ω



Notes: This figure presents the correlations among error terms $\epsilon_T, \epsilon_M, \epsilon_Y$. Properties of error terms are: *(i)* normally distributed with mean zero and variance one; *(ii)* $\epsilon_T \perp\!\!\!\perp \epsilon_Y$ and *(iii)* $\epsilon_M = \sqrt{\omega} \cdot \epsilon_T + \sqrt{(1-\omega)} \cdot \epsilon_Y$ where $\omega \in [0, 1]$. Parameters $\rho_{TM}, \rho_{MY}, \rho_{TY}$ stand for the correlations between (ϵ_T, ϵ_M), (ϵ_M, ϵ_Y), and (ϵ_T, ϵ_Y) respectively. Parameter $\rho_{TY|\epsilon_M}$ stands for the correlation between ϵ_T, ϵ_Y conditioned on error term ϵ_M.

It is instructive to investigate the bias generated by a misspecified model in which T, M are assumed to be exogenous, i.e. in which the mutual independence of $\epsilon_T, \epsilon_M, \epsilon_Y$ is wrongly assumed. Let the data be generated by equations (**??**)–(**??**), and the model coefficients be be normalized to equal 1, that is, $\beta_T^Z = \beta_M^T = \beta_Y^T = \beta_Y^M = 1$.The true parameters β_M^T, β_Y^T and β_Y^M are identified through equations (**??**)–(**??**). If the error terms $\epsilon_T, \epsilon_Y, \epsilon_M$ were wrongly assumed to be statistically independent, parameters $\beta_M^T, \beta_Y^T, \beta_Y^M$ could be estimated by OLS through the following equations:

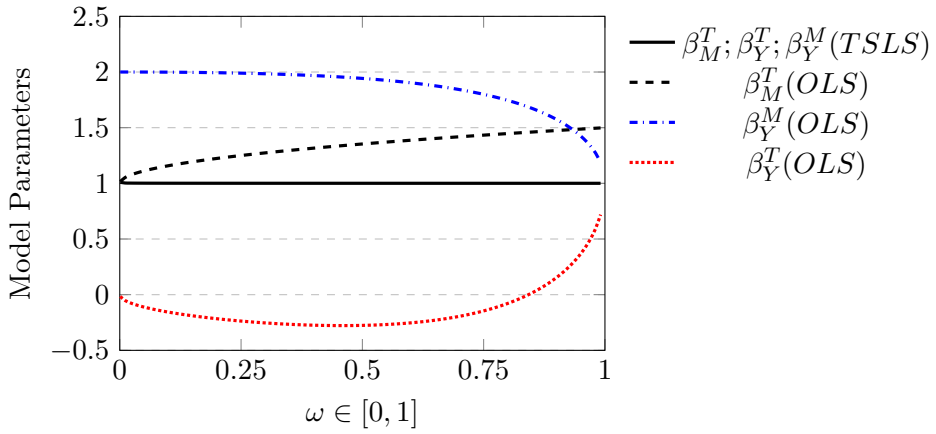
$$\text{OLS:} \quad \beta_M^T = \frac{\sigma_{TM}}{\sigma_{TY}}, \tag{11}$$

$$\text{OLS:} \quad \beta_Y^T = \frac{\sigma_{MM}\sigma_{TY} - \sigma_{TM}\sigma_{MY}}{\sigma_{MM}\sigma_{TT} - \sigma_{TM}^2}, \tag{12}$$

$$\text{OLS:} \quad \beta_Y^M = \frac{-\sigma_{TM}\sigma_{TY} + \sigma_{TT}\sigma_{MY}}{\sigma_{MM}\sigma_{TT} - \sigma_{TM}^2}. \tag{13}$$

Figure A2 displays the correct model parameters – evaluated by the 2SLS in (**??**)–(**??**) – and the OLS estimators identified by equations (11)–(13). While the true parameters are set to be 1, the OLS estimators may range from 0 to 2 depending on the error correlations. Since a high ω implies pronounced bias in the relation between T and M (a high ρ_{TM}), the OLS estimate of β_M^T diverges from the true value 1 as ω increases. By contrast, the OLS estimates of β_Y^T and β_Y^M converges to the true value 1.

Figure A2: Model Parameters by ω



Notes: The figure presents the model parameters $\beta_M^T, \beta_Y^T, \beta_Y^M$ computed under the right assumption that error correlate and under the mistaken assumption of no error correlation. If errors correlate according to **??**, then parameters β_M^T, β_Y^T and β_Y^M are identified by equations (**??**)–(**??**). Under the (wrong) assumption of no error correlation, the model parameters are identified by (11)–(13).

Appendix C Examining the Bounds Using Simulated Data

For the purposes of section **??**, it is instructive to simulate a model that describes the basic features of the estimation method under correlated error terms. Appendix C.1 defines the equations that generate the correlated error terms, and simulates the data. Appendix C.2 describes a procedure for how $\rho_{TY}(\kappa)$ can be identified as a function of κ, and how the other four parameters β_Y^M, β_Y^T, ρ_{MY}, $\sigma^2_{\epsilon_Y}$ can be in turn identified.

Appendix C.1 A Structure for Correlated Error Terms

The error terms are generated on the basis of a parameter $\omega \in [0,1]$ and four random variables $\xi_T, \xi_Y, \xi_M, \xi_E$ that are i.i.d. normally distributed with mean zero and variance 1, $\mathbf{N}(0,1)$. The error structure is defined by the following equations:

$$\epsilon_T = \sqrt{\omega} \cdot \xi_E + \sqrt{(1-\omega)} \cdot \xi_T; \tag{14}$$

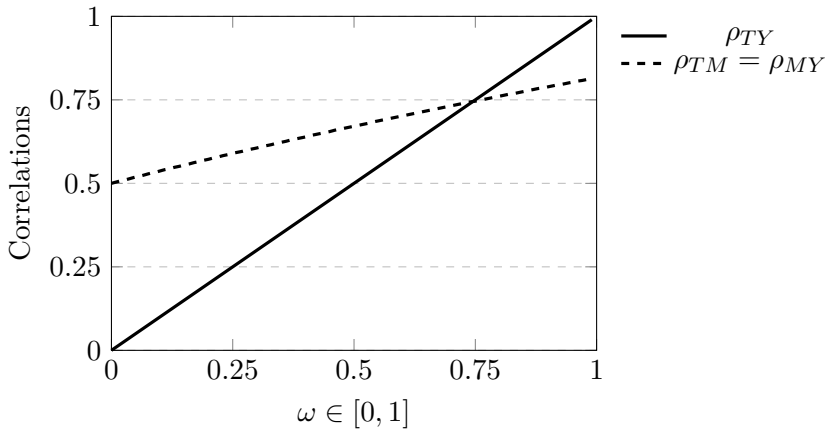
$$\epsilon_Y = \sqrt{\omega} \cdot \xi_E + \sqrt{(1-\omega)} \cdot \xi_Y; \tag{15}$$

$$\epsilon_M = \sqrt{0.5} \cdot \xi_M + \sqrt{0.5} \cdot \Big(\sqrt{0.5} \cdot \epsilon_T + \sqrt{(0.5)} \cdot \epsilon_Y \Big); \tag{16}$$

Note that the correlation between error variables ϵ_T, ϵ_Y is given by $\rho_{TY} = \omega \in [0,1]$. Equation (16) of error term ϵ_M is symmetric regarding errors ϵ_T and ϵ_Y. As a consequence, the correlations between ϵ_M, ϵ_T and ϵ_M, ϵ_Y are the same, that is, $\rho_{TM} = \rho_{MY}$. These correlations also depend on parameter ω. For instance, when $\omega = 0$ implies that $\rho_{TM} = \rho_{MY} = 0.5$ and $\omega = 1$ implies that $\rho_{TM} = \rho_{MY} = 0.815$. Figure A3 displays the relation between correlations $\rho_{YT}, \rho_{TM}, \rho_{MY}$ and parameter ω.[5]

[5] Equations (14)–(16) also imply that error terms $\epsilon_T, \epsilon_M, \epsilon_Y$ are normally distributed with mean zero and variance 1. Finally, error term ϵ_Z is set to be normally distributed with mean zero and variance 1 and to be statistically independent of error terms $\epsilon_T, \epsilon_M, \epsilon_Y$.

Figure A3: Correlations Among Error Terms by ω



Notes: This figure presents the correlations among error terms

$$\epsilon_T = \sqrt{\omega} \cdot \xi_E + \sqrt{(1-\omega)} \cdot \xi_T;$$
$$\epsilon_Y = \sqrt{\omega} \cdot \xi_E + \sqrt{(1-\omega)} \cdot \xi_Y;$$
$$\epsilon_M = \sqrt{0.5} \cdot \xi_M + \sqrt{0.5} \cdot \left(\sqrt{0.5} \cdot \epsilon_T + \sqrt{(0.5)} \cdot \epsilon_Y\right);$$

where $\omega \in [0,1]$ and $\xi_T, \xi_Y, \xi_M, \xi_E$ are i.i.d normally distributed random variables with mean zero and variance 1.

Appendix C.2 Bounding the Model Parameters

The subsequent identification follows the same steps utilized in in section **??**. Model identification relies on the matrix equation $\widetilde{\Sigma}_{\mathbf{X}} = \Sigma_{\boldsymbol{\epsilon}}$, where $\widetilde{\Sigma}_{\mathbf{X}} \equiv (\mathbf{I} - \boldsymbol{\Psi}) \Sigma_{\mathbf{X}} (\mathbf{I} - \boldsymbol{\Psi})'$, as in (**??**). This yields ten linear equalities given by $\widetilde{\Sigma}_{\mathbf{X}}[i,j] = \Sigma_{\boldsymbol{\epsilon}}[i,j]$ for $i \leq j; i,j \in \{1,2,3,4\}$. The independence relation $Z \perp\!\!\!\perp (\epsilon_T, \epsilon_M)$ implies that $\Sigma_{\boldsymbol{\epsilon}}[1,2] = \Sigma_{\boldsymbol{\epsilon}}[1,3] = 0$. Thereby, the equalities $\widetilde{\Sigma}_{\mathbf{X}}[1,2] = 0$ and $\widetilde{\Sigma}_{\mathbf{X}}[1,3] = 0$ as in (**??**)–(**??**) still hold. As a consequence, the coefficients β_T^Z, β_M^T remain unchanged and are still identified by $\beta_T^Z = \frac{\sigma_{ZT}}{\sigma_{ZZ}}$ and $\beta_M^T = \frac{\sigma_{ZM}}{\sigma_{ZT}}$. The coefficients β_T^Z and β_M^T can still be evaluated by the OLS regression in (**??**) and the 2SLS in (**??**)–(**??**), respectively. The coefficients β_T^Z, β_M^T refer to *Model I* in Table 1. The model is not altered by the causal relation between T and Y, and therefore β_T^Z, β_M^T are not affected by relaxing $\rho_{TY} \neq 0$.

Error variances are identified by the diagonal of the matrix equality $\widetilde{\Sigma}_{\mathbf{X}} = \Sigma_{\boldsymbol{\epsilon}}$. The equality $\widetilde{\Sigma}_{\mathbf{X}}[1,1] = \Sigma_{\mathbf{e}}[1,1]$ implies that $\sigma_{ZZ} = \sigma^2_{\epsilon_Z}$. Furthermore, the identification of β_T^Z and the observed

covariances enable the identification of error variance $\sigma^2_{\epsilon_T}$ by the following equation:

$$\widetilde{\Sigma}_{\mathbf{X}}[2,2] = \Sigma_{\mathbf{e}}[2,2] \quad \Rightarrow \quad \left(\sigma_{TT} - \beta_T^Z \sigma_{ZT}\right) - \beta_T^Z \left(\sigma_{ZT} - \beta_T^Z \sigma_{ZZ}\right) \quad = \quad \sigma^2_{\epsilon_T}. \tag{17}$$

In addition, the identification of β_M^T enables the identification of $\sigma^2_{\epsilon_M}$ by the following equation:

$$\widetilde{\Sigma}_{\mathbf{X}}[3,3] = \Sigma_{\mathbf{e}}[3,3] \quad \Rightarrow \quad \left(\sigma_{MM} - \beta_M^T \sigma_{TM}\right) - \beta_M^T \left(\sigma_{TM} - \beta_M^T \sigma_{TT}\right) \quad = \quad \sigma^2_{\epsilon_M}. \tag{18}$$

The parameters β_T^Z, β_M^T and variances $\sigma^2_{\epsilon_T}, \sigma^2_{\epsilon_M}$ enable the identification of correlation ρ_{TM} via the following equation:

$$\widetilde{\Sigma}_{\mathbf{X}}[2,3] = \Sigma_{\mathbf{e}}[2,3] \quad \Rightarrow \quad \sigma_{TM} - \beta_T^Z \sigma_{ZM} - \beta_M^T(\sigma_{TT} - \beta_T^Z \sigma_{ZT}) \quad = \quad \rho_{TM}\sigma_{\epsilon_T}\sigma_{\epsilon_M}. \tag{19}$$

We are left with four equalities, $\widetilde{\Sigma}_{\mathbf{X}}[i,j] = \Sigma_{\mathbf{e}}[i,j]$ such that $(i,j) \in \{(1,4),(2,4),(4,4),(3,4)\}$, and five parameters, $\beta_Y^M, \beta_Y^T, \rho_{TY}, \rho_{MY}, \sigma^2_{\epsilon_Y}$, that are not point-identified.

The best approach to examine the identification of the five model parameters that are not point-identified is to express all five parameters $\beta_Y^M, \beta_Y^T, \rho_{TY}, \rho_{MY}, \sigma^2_{\epsilon_Y}$, in terms of the product $\rho_{TY} \cdot \sigma_{\epsilon_Y}$, which we label as an auxiliary variable $\kappa \equiv \rho_{TY} \cdot \sigma_{\epsilon_Y}$. If κ were known, we could identify all model parameters. Specifically, let $\beta_Y^M(\kappa)$, $\beta_Y^T(\kappa)$, $\rho_{TY}(\kappa)$, $\rho_{MY}(\kappa)$, $\sigma^2_{\epsilon_Y}(\kappa)$ denote the values that the model parameters would take for a given value of κ. Let $\beta_Y^M(\kappa)$, $\beta_Y^T(\kappa)$, $\rho_{TY}(\kappa)$, $\rho_{MY}(\kappa)$, $\sigma^2_{\epsilon_Y}(\kappa)$ denote the values that the model parameters would take for a given value of κ. These parameters can be identified by the following procedure:

1. The equalities $\widetilde{\Sigma}_{\mathbf{X}}[1,4] = \Sigma_{\mathbf{e}}[1,4]$, and $\widetilde{\Sigma}_{\mathbf{X}}[2,4] = \Sigma_{\mathbf{e}}[2,4]$ generate the following equations:

$$\sigma_{ZY} - \beta_Y^M \sigma_{ZM} - \beta_Y^T \sigma_{ZT} = 0, \tag{20}$$

$$\sigma_{TY} - \beta_Y^M \sigma_{TM} - \beta_Y^T \sigma_{TT} = \underbrace{\rho_{TY}\sigma_{\epsilon_Y}}_{\kappa} \sigma_{\epsilon_T}. \tag{21}$$

Given a value of κ, the coefficients $\beta_Y^M(\kappa), \beta_Y^T(\kappa)$ can be obtained by the following formula:

$$\begin{bmatrix} \beta_Y^M(\kappa) \\ \beta_Y^T(\kappa) \end{bmatrix} = \left(\mathbf{B}'\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \cdot \left(\mathbf{B}'\mathbf{A}^{-1}\mathbf{C}\right), \tag{22}$$

$$\text{where } \mathbf{A} = \begin{bmatrix} \sigma_{ZZ} & \sigma_{ZT} \\ \sigma'_{ZT} & \sigma_{TT} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \sigma_{ZM} & \sigma_{ZT} \\ \sigma'_{TM} & \sigma_{TT} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \sigma_{ZY} \\ \sigma_{TY} - \kappa \cdot \sigma_{\epsilon_T} \end{bmatrix}. \tag{23}$$

2. The parameters $\beta_Y^M(\kappa), \beta_Y^T(\kappa)$ enable the evaluation of error variance $\sigma^2_{\epsilon_Y}(\kappa)$ via equality $\widetilde{\Sigma}_{\mathbf{X}}[4,4] = \Sigma_{\mathbf{e}}[4,4]$ in (**??**). Namely:

$$\sigma^2_{\epsilon_Y}(\kappa) = \begin{pmatrix} 1 \\ -\beta_Y^M(\kappa) \\ -\beta_Y^T(\kappa) \end{pmatrix}' \begin{bmatrix} \sigma_{YY} & \sigma_{MY} & \sigma_{TY} \\ \sigma_{MY} & \sigma_{MM} & \sigma_{TM} \\ \sigma_{TY} & \sigma_{TM} & \sigma_{TT} \end{bmatrix} \begin{pmatrix} 1 \\ -\beta_Y^M(\kappa) \\ -\beta_Y^T(\kappa) \end{pmatrix}. \tag{24}$$

3. The evaluation of $\sigma_{\epsilon_Y}(\kappa)$ in addition to $\beta_Y^T(\kappa), \beta_Y^M(\kappa)$ as well as $\beta_M^T, \sigma_{\epsilon_M}$ enables the identification of the correlation $\rho_{MY}(\kappa)$ via the equality $\widetilde{\Sigma}_{\mathbf{X}}[3,4] = \Sigma_{\mathbf{e}}[3,4]$ in (**??**). Namely:

$$\rho_{MY}(\kappa) \quad = \quad \begin{pmatrix} 1 \\ -\beta_Y^M(\kappa) \\ -\beta_Y^T(\kappa) \end{pmatrix}' \begin{bmatrix} \sigma_{MY} & \sigma_{TY} \\ \sigma_{MM} & \sigma_{TM} \\ \sigma_{TM} & \sigma_{TT} \end{bmatrix} \begin{pmatrix} \left(\sigma_{\epsilon_M}\sigma_{\epsilon_Y}(\kappa)\right)^{-1} \\ \frac{-\beta_M^T}{\sigma_{\epsilon_M}(\kappa)\sigma_{\epsilon_Y}(\kappa)} \end{pmatrix}. \tag{25}$$

4. Finally, the correlation $\rho_{TY}(\kappa)$ can be identified by the ratio $\rho_{TY}(\kappa) = \kappa/\sigma_{\epsilon_Y}(\kappa)$.

The identification of model parameters is thus anchored on the variable κ. The variable κ^* is unknown, but we can bound it on the interval $\kappa \in [\kappa_0, \kappa_1]$ by imposing some simple model restrictions on the the covariance structure of the data: (i) Parameter $\rho_{TY}(\kappa)$ stands for the correlation between ϵ_T and ϵ_Y. Thus, we can delimit the values of κ such that $|\rho_{TY}(\kappa)| \leq 1$. Likewise, parameter $\rho_{MY}(\kappa)$ stands for the correlation between ϵ_M and ϵ_Y, such that $|\rho_{MY}(\kappa)| \leq 1$ must hold. (ii) Let $\Sigma_{\mathbf{e}}(\kappa)$ denote the identified error covariance matrix for a value of κ. $\Sigma_{\mathbf{e}}(\kappa)$ must be positive-definite matrix, that is to say that its eigenvalues must be strictly positive. Thus the values of κ must be such that the root-solutions of the λ-polynomial generated by the determinant $\det\left(\Sigma_{\mathbf{e}}(\kappa) - \lambda\mathbf{I}\right) = 0$ are strictly positive.[6] (iii) The outcome error variance $\sigma^2_{\epsilon_Y}$ is smaller than or equal to the variance of the outcome σ_{YY} itself, i.e. $\sigma_{YY} \geq \sigma^2_{\epsilon_Y}$. Because κ is defined as $\kappa \equiv \rho_{TY} \cdot \sigma_{\epsilon_Y}$, it must therefore lie in the interval $\kappa \in [-\sqrt{\sigma_{\epsilon_Y}}, \sqrt{\sigma_{\epsilon_Y}}]$. (iv) Lastly, we can add the restriction $0 \leq \rho_{MY}(\kappa) \leq 1$ on the model correlation.

We evaluate the interval $[\kappa_0, \kappa_1]$ that complies with all these model restrictions. With this interval, we can bound ρ_{TY}, and then bound the other four parameters as functions of ρ_{TY}. We can thus generate the bounds for the direct effect $DE = \beta_Y^T$; the indirect effect $IE = \beta_M^T \cdot \beta_Y^M$; the total effect $TE = \beta_Y^T + \beta_M^T \cdot \beta_Y^M$; the share of the total effect that is mediated by the indirect effect.

[6] In our notation, $\mathbf{I}$ stands for the identity matrix of dimension 4.

Online Appendix

to

"Mediation Analysis in IV Settings With a Single Instrument"

Online Appendix A The Mediation Model with No Confounders

It is illustrate how mediation works in the setting above in a simple randomized control trial (RCT), as in the seminal work of Robins and Greenland (1992), and using the language of counterfactual variables. In a mediation model, the causal effect of T on Y decomposes into the *direct*, and the *indirect* effects. The total effect TE stands for the average causal effect of T on Y. The direct effect DE stands for the causal effect of T on Y that is not generated by changes in M. The indirect effect IE is the causal effect of T on Y induced by the change in the distribution of the mediator M. Let the treatment assignment take values in $\text{supp}(T) = \{t_0, t_1\}$, where t_0 indicates the control group and t_1 indicates the treatment group. Let $F_{M(t)}(m)$ denote the cumulative density function (CDF) of the counterfactual mediator $M(t)$ conditional on the assignment $t \in \{t_0, t_1\}$. The total effect TE is the expected difference between counterfactual outcome Y when T is fixed at t_1 and t_0. The direct effect $DE(t)$ evaluates the expected difference of counterfactual outcomes between treated (t_1) and control (t_0) group holding the distribution of the mediator fixed at $M(t)$. The indirect effect $IE(t)$ evaluates the expected value of the the difference between counterfactual outcomes $Y(t, m)$ when the distribution of the mediator m varies between treated $M(t_1)$ and control $M(t_0)$ while holding the t-input fixed. Formally, the other three causal effects (TE, DE, IE) are defined as follows:

$$\begin{array}{lcccc} TE & = & E(Y(t_1) - Y(t_0)) & \equiv & E(Y(t_1, M(t_1)) - Y(t_0, M(t_0))) \\ DE(t) & = & E(Y(t_1, M(t)) - Y(t_0, M(t))) & \equiv & \int E\Big(Y(t_1, m) - Y(t_0, m)\Big) dF_{M(t)}(m) \\ IE(t) & = & E(Y(t, M(t_1)) - Y(t, M(t_0))) & \equiv & \int E\Big(Y(t, m)\Big) \Big[dF_{M(t_1)}(m) - dF_{M(t_0)}(m)\Big] \end{array}$$

Robins and Greenland's (1992) main contribution is to show that the total effect of T on Y can be decomposed as the sum of the effect of T on Y that is mediated by M (the indirect effect) and the causal effect of T on Y that is not mediated by M (the direct effect).[7] Equations (26)–(27) express the total effect as the sum of direct and indirect effects:[8]

$$\begin{aligned} TE &= E(Y(t_1, M(t_1)) - Y_i(t_0, M(t_0))) \\ &= \Big(E(Y(t_1, M(t_1))) - E(Y(t_0, M(t_1)))\Big) + \Big(E(Y(t_0, M(t_1)) - Y_i(t_0, M(t_0)))\Big) = DE(t_1) + IE(t_0) \qquad (26) \\ &= \Big(E(Y(t_1, M(t_1))) - E(Y(t_1, M(t_0)))\Big) + \Big(E(Y(t_1, M(t_0)) - Y_i(t_0, M(t_0)))\Big) = IE(t_1) + DE(t_0). \qquad (27) \end{aligned}$$

Robins and Greenland's (1992) decomposition can be extended to the problem we examine by allowing T to be a continuous variable, in which case the decomposition is obtained by the total differentiation of the counterfactual outcome:

$$\underbrace{\frac{dE(Y(t))}{dt}}_{\text{Total Effect}} = \underbrace{\frac{\partial E(Y(t,m))}{\partial t}}_{\text{Direct Effect}} + \underbrace{\frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt}}_{\text{Indirect Effect}}. \qquad (28)$$

[7] Pearl (2011) makes a distinction between natural (or "descriptive") direct and indirect effects and controlled (or "prescriptive") direct effects.

[8] A large literature on mediation analysis relies on the Sequential Ignorability Assumption **A-1** of Imai et al. (2010) to identify mediation effects. We discuss this assumption in Online Appendix B. See Frölich and Huber (2017) for a recent review of the mediation literature.

Identification of the total, direct, and indirect effects hinges on the dependence relation among the error terms $\epsilon_T, \epsilon_M, \epsilon_Y$ in (**??**)–(**??**). Suppose that the error terms ϵ_T and ϵ_M are statistically independent, i.e. $\epsilon_T \perp\!\!\!\perp \epsilon_M$. This means there are no unobserved variables that jointly cause T and M. In this case, T is exogenous with respect to M, as in an RCT. It is easy to show that the independence condition $M(t) \perp\!\!\!\perp T$ holds and the expected value of counterfactual variable $M(t)$ is identified by the conditional expectation $E(M(t)) = E(M|T = t)$. In addition, if error terms (ϵ_T, ϵ_M) and ϵ_Y were statistically independent, then the independence conditions $\big(Y(t), Y(t,m)\big) \perp\!\!\!\perp T$ and $Y(t,m) \perp\!\!\!\perp M$ would also hold. This means that variables T and M are exogenous with respect to Y and that the expected value of counterfactual variables $E(Y(t))$ and $E(Y(t,m))$ would be identified by conditional expectations of observed variables $E(Y(t)) = E(Y|T = t)$ and $E(Y(t,m)) = E(Y|T = t, M = m)$, respectively.

Online Appendix B The Sequential Ignorability Assumption

A large literature on mediation analysis relies on the Sequential Ignorability Assumption **A-1** of Imai et al. (2010) to identify mediation effects.

Assumption A-1. Sequential Ignorability (Imai et al., 2010):

$$\big(Y(t',m), M(t)\big) \perp\!\!\!\perp T|X \tag{29}$$

$$Y(t',m) \perp\!\!\!\perp M(t)|(T,X), \tag{30}$$

where X denotes pre-intervention variables that are not caused by T, M and Y such that $0 < P(T=t|X) < 1$ and $0 < P(M(t)=m|T=t,X) < 1$ holds for all $x \in \text{supp}(X)$ and $m \in \text{supp}(M)$.

Under Sequential Ignorability **A-1**, it is easy to show that the distributions of counterfactual variables are identified by $P(Y(t,m)|X) = P(Y|X,T=t,M=m)$ and $P(M(t)|X) = P(M|X,T=t)$ and thereby the mediating causal effects can be expressed as:

$$ADE(t) = \int \left(\big(E(Y|T=t_1,M=m,X=x) - E(Y|T=t_0,M=m,X=x,X=x)\big) dF_{M|T=t,X=x}(m) \right) dF_X(x) \tag{31}$$

$$AIE(t) = \int \left(E(Y|T=t,M=m,X=x) \Big[dF_{M|T=t_1,X=x}(m) - dF_{M|T=t_0,X=x}(m) \Big] \right) dF_X(x). \tag{32}$$

Imai, Tingley, Keele and Yamamoto offer a substantial line of research that explores the identifying properties of Sequential Ignorability Assumption **A-1**. See Imai et al. (2011a) for a comprehensive discussion of the benefits and limitations of the sequential ignorability assumption.

The main critics of Sequential Ignorability **A-1** is that it does not hold under the presence of either *Confounders* or *Unobserved Mediators* (Heckman and Pinto, 2015).

The independence relation (29) assumes that T is exogenous conditioned on X. There exists no unobserved variable that causes T and Y or T and M. For instance, the Sequential Ignorability **A-1** holds for the model defined in (**??**) because:

$$\big(\epsilon_Y, \epsilon_M\big) \perp\!\!\!\perp \epsilon_T \Rightarrow \qquad \big(f_Y(t',m,\epsilon_Y), f_M(t,\epsilon_M)\big) \perp\!\!\!\perp f_T(\epsilon_T) \Rightarrow \qquad \big(Y(t',m), M(t)\big) \perp\!\!\!\perp T. \tag{33}$$

$$\epsilon_Y \perp\!\!\!\perp \epsilon_M|\epsilon_T \Rightarrow \qquad f_Y(t',m,\epsilon_Y) \perp\!\!\!\perp f_M(t,\epsilon_M)|f_T(\epsilon_T) \Rightarrow \qquad Y(t',m) \perp\!\!\!\perp M(t)|T, \tag{34}$$

where the initial independence relation in (33) and (34) comes from the independence of error terms.

This assumption is expected to hold in experimental data when treatment T is randomly assigned. The independence relation (30) assumes that M is exogenous conditioned on X and T. It assumes that no confounding variable causing M and Y. Sequential Ignorability **A-1** is an extension of the Ignorability Assumption of Rosenbaum and Rubin (1983) that also assumes that a treatment T is exogenous when conditioned on pre-treatment variables. Petersen et al. (2006); Robins (2003); Rubin (2004) state similar identifying criteria that assume no confounding variables. Those assumptions are not testable.

Sequential Ignorability **A-1** assumes that: (1) the confounding variable V is observed, that is, the pre-treatment variables X; and (2) that there is no unobserved mediator U. This assumption is unappealing for many because it solves the identification problem generated by confounding variables by assuming that those do not exist (Heckman, 2008).

Consider a change in the treatment variable T denoted by $\Delta(t) = t_1 - t_0$. The Direct and

indirect effects can be expressed by:

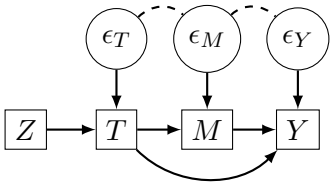
$$
\begin{aligned}
ADE(t') &= \Big(\lambda_{YT} \cdot t_1 + \lambda_{YM} \cdot E(M(t'))\Big) - \Big(\lambda_{YT} \cdot t_0 + \lambda_{YM} \cdot E(M(t'))\Big) \\
\therefore ADE &= \lambda_{YT} \cdot \Delta(t) \qquad (35) \\
\text{and } AIE(t') &= \Big(\lambda_{YT} \cdot t' + \lambda_{YM} \cdot E(M(t_1))\Big) - \Big(\lambda_{YT} \cdot t' + \lambda_{YM} \cdot E(M(t_0))\Big) \\
&= \Big(\lambda_{YT} \cdot t' + \lambda_{YM}\lambda_M \cdot t_1\Big) - \Big(\lambda_{YT} \cdot t' + \lambda_{YM}\lambda_M \cdot t_0\Big) \\
\therefore AIE &= \lambda_{YM} \cdot \lambda_M \cdot \Delta(t) \qquad (36)
\end{aligned}
$$

Online Appendix C The IV Mediate Model under Assumption ??

Panel A of Table Online Appendix Table 1 represents the causal relations of the mediation model (**??**)–(**??**) as a directed acyclic graph (DAG). Squares represent observed variables, while circles denote unobserved variables. Causal relations are denoted by solid lines while the dependence structure among error variables is depicted by dashed lines. Table Online Appendix Table 1 is a version of *Model III* in Table 1 that accounts for the statistical dependence among error terms in **??**.

Table Online Appendix Table 1: The Mediation Model with IV

A. DAG Representation



B. Model Equations

Treatment variable:	$T = f_T(Z, \epsilon_T)$
Observed mediator:	$M = f_M(T, \epsilon_M)$
Outcome:	$Y = f_Y(T, M, \epsilon_Y)$
where:	$\epsilon_T \not\perp\!\!\!\perp \epsilon_M$, $\epsilon_M \not\perp\!\!\!\perp \epsilon_Y$, $\epsilon_T \not\perp\!\!\!\perp \epsilon_Y \vert \epsilon_M$
and:	$Z \perp\!\!\!\perp \left(\epsilon_T, \epsilon_M, \epsilon_Y\right), \epsilon_T \perp\!\!\!\perp \epsilon_Y$

Online Appendix D Identification of Causal Parameters

When we additionally allow for an unobserved mediator U that is caused by T and causes both M and Y (see **Remark ??**), the linear mediation model we investigate can be fully described by the following equations:

$$\begin{aligned}
\text{Instrumental Variable } Z &= \epsilon_Z, && (37)\\
\text{Treatment } T &= \xi_Z \cdot Z + \xi_V \cdot V_T + \epsilon_T, && (38)\\
\text{Unobserved Mediator } U &= \zeta_T \cdot T + \epsilon_U, && (39)\\
\text{Observed Mediator } M &= \varphi_T \cdot T + \varphi_U \cdot U + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M, && (40)\\
\text{Outcome } Y &= \beta_T \cdot T + \beta_M \cdot M + \beta_U \cdot U + \beta_V \cdot V_Y + \epsilon_Y, && (41)\\
\text{Exogenous Variables} \quad & Z, V_T, V_M, \epsilon_Z, \epsilon_T, \epsilon_U, \epsilon_M, \epsilon_Y \text{ are statistically independent variables}, && (42)\\
\text{Scalar Coefficients} \quad & \xi_Z, \xi_V, \zeta_T, \varphi_T, \varphi_U, \delta_Y, \delta_T, \beta_T, \beta_M, \beta_U, \beta_V && (43)\\
\text{Unobserved Variables} \quad & V_T, V_M, U, \epsilon_Z, \epsilon_T, \epsilon_U, \epsilon_M, \epsilon_Y. && (44)
\end{aligned}$$

We assume that all variables have mean zero. This assumption does not incur in less of generality, but simplify notation as intercepts can be suppressed.

We first eliminate the unobserved mediator U from Equations (40)–(41) by iterated substitution. Equations (41)–(41) are then expressed as:

$$\begin{aligned}
M &= (\varphi_T + \varphi_U\zeta_T) \cdot T + \varphi_U \cdot \epsilon_U + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M, && (45)\\
Y &= (\beta_T + \beta_U\zeta_T) \cdot T + \beta_M \cdot M + \beta_U \cdot \epsilon_U + \beta_V \cdot V_Y + \epsilon_Y. && (46)
\end{aligned}$$

We use the following transformation of parameters to save on notation:

$$\begin{aligned}
\widetilde{\varphi}_T &= \varphi_T + \varphi_U\zeta_T, && (47)\\
\widetilde{\beta}_T &= \beta_T + \beta_U\zeta_T, && (48)\\
\widetilde{U} &= \epsilon_U. && (49)
\end{aligned}$$

We use equations (45)–(49) to simplify Model (37)–(41) into the following equations:

$$\begin{aligned}
\text{Instrumental Variable } Z &= \epsilon_Z, && (50)\\
\text{Treatment } T &= \xi_Z \cdot Z + \xi_V \cdot V_T + \epsilon_T, && (51)\\
\text{Observed Mediator } M &= \widetilde{\varphi}_T \cdot T + \varphi_U \cdot \widetilde{U} + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M, && (52)\\
\text{Outcome } Y &= \widetilde{\beta}_T \cdot T + \beta_M \cdot M + \beta_U \cdot \widetilde{U} + \beta_V \cdot V_Y + \epsilon_Y. && (53)
\end{aligned}$$

In this linear model, the counterfactual outcomes $M(t), Y(t), Y(m), Y(m,t)$ are given by:

$$M(t) = \widetilde{\varphi}_T \cdot t + \varphi_U \cdot \widetilde{U} + \delta_Y \cdot V_Y + \delta_T \cdot V_T + \epsilon_M, \tag{54}$$

$$Y(m) = \widetilde{\beta}_T \cdot T + \beta_M \cdot m + \beta_U \cdot \widetilde{U} + \beta_V \cdot V_Y + \epsilon_Y. \tag{55}$$

$$Y(t,m) = \widetilde{\beta}_T \cdot t + \beta_M \cdot m + \beta_U \cdot \widetilde{U} + \beta_V \cdot V_Y + \epsilon_Y. \tag{56}$$

$$\begin{aligned} Y(t) &= \widetilde{\beta}_T \cdot t + \beta_M \cdot M(t) + \beta_U \cdot \widetilde{U} + \beta_V \cdot V_Y + \epsilon_Y. \\ &= (\widetilde{\beta}_T + \beta_M \widetilde{\varphi}_T) \cdot t + (\beta_U + \beta_M \varphi_U) \cdot \widetilde{U} + (\beta_V + \beta_M \delta_Y) \cdot V_Y + \beta_M \delta_T \cdot V_T + \beta_M \cdot \epsilon_M + \epsilon_Y. \end{aligned} \tag{57}$$

We claim that the coefficients associated with unobserved variables $V_T, \widetilde{U}, V_Y$ may only be identified up a linear transformation. Consider the coefficients δ_T, β_V that multiply the unobserved variable V_T in Equations (51) and (52) respectively. Suppose a linear transformation that multiplies V_T by a constant $\kappa \neq 0$. The model would remain the same if coefficients δ_T, β_V were divided by the same constant κ. This is a typical fact in the literature of linear factor models. We solve this non-identification problem by impose that each unobserved variable $V_T, \widetilde{U}, V_Y$ has unit variance:

$$\text{var}(V_T) = \text{var}(\widetilde{U}) = \text{var}(V_Y) = 1. \tag{58}$$

Assumption (58) is typically termed as *anchoring* of unobserved factors in the literature of factor analysis. This assumption does not incur in any loss of generality for the identification of direct, indirect or total causal effects of T (and M) on Y as expressed in the following section.

Online Appendix D.1 Defining Causal Parameters

The literature of mediation analysis term relevant causal parameters as:

- Total Effect of T on Y, that is, $\frac{dE(Y(t))}{dt}$.
- Direct Effect of T on Y, that is $\frac{\partial E(Y(t,m))}{\partial t}$.
- Effect of M on Y, that is, $\frac{dE(Y(m))}{dm}$.
- Effect of T on M, that is, $\frac{dE(M(t))}{dt}$.
- Indirect Effect of T on Y, that is $\frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt}$.

According to the counterfactual variables in (54)–(57), these causal effects are given by:

$$\text{Total Effect of } T \text{ on } Y: \frac{dE(Y(t))}{dt} = \widetilde{\varphi}_T \cdot \beta_M + \widetilde{\beta}_T. \tag{59}$$

$$\text{Direct Effect of } T \text{ on } Y: \frac{\partial E(Y(t,m))}{\partial t} = \widetilde{\beta}_T. \tag{60}$$

$$\text{Effect of } M \text{ on } Y: \frac{dE(Y(m))}{dm} = \beta_M. \tag{61}$$

$$\text{Effect of } T \text{ on } M: \frac{dE(M(t))}{dt} = \widetilde{\varphi}_T. \tag{62}$$

$$\text{Indirect Effect of } T \text{ on } Y: \frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt} = \beta_M \cdot \widetilde{\varphi}_T. \tag{63}$$

Online Appendix D.2 Identifying Equations

Model (50)–(53) can be conveniently expressed in matrix notation. In Equation (64) we define $\mathbf{X} = [Z, T, M, Y]'$ as the vector of observed variables, $\mathbf{V} = [V_T, V_Y, \widetilde{U}]'$ as the vector of unobserved confounding variables, and $\boldsymbol{\epsilon} = [\epsilon_Z, \epsilon_T, \epsilon_M, \epsilon_Y]'$ as the vector of exogenous error terms. According to (42), the random vectors $\mathbf{V}$ and $\boldsymbol{\epsilon}$ are independent, that is, $\mathbf{V} \perp\!\!\!\perp \boldsymbol{\epsilon}$. We use $\mathbf{K}$ in (64) for the matrix of parameters that multiply $\mathbf{X}$ and $\mathbf{A}$ for the matrix of parameters that multiply $\mathbf{V}$.

$$\mathbf{X} = \begin{pmatrix} Z \\ T \\ M \\ Y \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} V_T \\ V_Y \\ \widetilde{U} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_Z \\ \epsilon_T \\ \epsilon_M \\ \epsilon_Y \end{pmatrix}, \quad \mathbf{K} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \xi_Z & 0 & 0 & 0 \\ 0 & \widetilde{\varphi}_T & 0 & 0 \\ 0 & \widetilde{\beta}_T & \beta_M & 0 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ \xi_V & 0 & 0 \\ \delta_Y & \delta_Y & \varphi_U \\ 0 & \beta_V & \beta_U \end{bmatrix}. \tag{64}$$

Using the notation in (64), we can express the linear system (50)–(53) as following:

$$\underbrace{\begin{pmatrix} Z \\ T \\ M \\ Y \end{pmatrix}}_{\mathbf{X}} = \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ \xi_Z & 0 & 0 & 0 \\ 0 & \widetilde{\varphi}_T & 0 & 0 \\ 0 & \widetilde{\beta}_T & \beta_M & 0 \end{bmatrix}}_{\mathbf{K}} \cdot \underbrace{\begin{pmatrix} Z \\ T \\ M \\ Y \end{pmatrix}}_{\mathbf{X}} + \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ \xi_V & 0 & 0 \\ \delta_T & \delta_Y & \varphi_U \\ 0 & \beta_V & \beta_U \end{bmatrix}}_{\mathbf{A}} \cdot \underbrace{\begin{pmatrix} V_T \\ V_Y \\ \widetilde{U} \end{pmatrix}}_{\mathbf{V}} + \underbrace{\begin{pmatrix} \epsilon_Z \\ \epsilon_T \\ \epsilon_M \\ \epsilon_Y \end{pmatrix}}_{\boldsymbol{\epsilon}}, \tag{65}$$

$$\mathbf{X} = \mathbf{K} \cdot \mathbf{X} + \mathbf{A} \cdot \mathbf{V} + \boldsymbol{\epsilon}. \tag{66}$$

The coefficients in matrices $\mathbf{K}, \mathbf{A}$ are identified through the covariance matrices of observed variables. We use $\Sigma_{\mathbf{X}} = \text{cov}(\mathbf{X}, \mathbf{X})$ for the covariance matrix of observed variables $\mathbf{X}$, and $\Sigma_{\boldsymbol{\epsilon}} = \text{cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})$ for the vector of error terms $\boldsymbol{\epsilon}$. $\Sigma_{\boldsymbol{\epsilon}}$ is a diagonal matrix due to statistical independence of error terms. We also use $\Sigma_{\mathbf{V}} = \text{cov}(\mathbf{V}, \mathbf{V})$ for the covariance of unobserved variables $\mathbf{V}$. The unobserved variables in $\mathbf{V}$ are statistically independent and have unit variance (58), thus $\Sigma_{\mathbf{V}} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. Moreover, $\mathbf{V} \perp\!\!\!\perp \boldsymbol{\epsilon}$ implies that $\text{cov}(\mathbf{V}, \boldsymbol{\epsilon}) = \mathbf{0}$, where $\mathbf{0}$ is a matrix of elements zero.

Equation (69) determines the relation between the covariance matrices of observed and unobserved variables:

$$\begin{aligned} \mathbf{X} = \mathbf{K} \cdot \mathbf{X} + \mathbf{A} \cdot \mathbf{V} + \boldsymbol{\epsilon} &\Rightarrow (\mathbf{K} - \mathbf{I})\, \mathbf{X} = \mathbf{A} \cdot \mathbf{V} + \boldsymbol{\epsilon}, && (67) \\ &\Rightarrow (\mathbf{K} - \mathbf{I})\, \Sigma_{\mathbf{X}}\, (\mathbf{K} - \mathbf{I})' = \mathbf{A} \Sigma_{\mathbf{V}} \mathbf{A}' + \Sigma_{\boldsymbol{\epsilon}}, && (68) \\ &\Rightarrow (\mathbf{K} - \mathbf{I})\, \Sigma_{\mathbf{X}}\, (\mathbf{K} - \mathbf{I})' = \mathbf{A}\mathbf{A}' + \Sigma_{\boldsymbol{\epsilon}}, && (69) \end{aligned}$$

where the second equation is due to $\mathbf{V} \perp\!\!\!\perp \boldsymbol{\epsilon}$ and the third equations comes from $\Sigma_{\mathbf{V}} = \mathbf{I}$.

Equation (69) generates ten equalities. Four equalities are due to the diagonal of the covariance matrices $(\mathbf{K} - \mathbf{I})\, \Sigma_{\mathbf{X}}\, (\mathbf{K} - \mathbf{I})'$ and $\mathbf{A}\mathbf{A}' + \Sigma_{\boldsymbol{\epsilon}}$ in (69). The remaining six equalities from the off-diagonal relation of the covariance matrices in (69).

The diagonal elements of $\Sigma_{\boldsymbol{\epsilon}}$ are the variances of the error terms $\epsilon_Z, \epsilon_T, \epsilon_M, \epsilon_Y$. Thereby each diagonal equation generated by (69) adds one unobserved term to the system of quadratic equations. The point-identification of the model coefficients arises from the six off-diagonal equations generated

by (69). Those are listed below:

$$\operatorname{cov}(Z,T) - \operatorname{cov}(Z,Z)\cdot\xi_Z = 0 \tag{70}$$
$$\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\cdot\widetilde{\varphi}_T = 0 \tag{71}$$
$$\operatorname{cov}(Z,Y) - \operatorname{cov}(Z,M)\cdot\beta_M - \operatorname{cov}(Z,T)\cdot\widetilde{\beta}_T = 0 \tag{72}$$
$$\operatorname{cov}(T,Y) - \operatorname{cov}(T,T)\cdot\widetilde{\beta}_T - \operatorname{cov}(T,M)\cdot\beta_M = 0 \tag{73}$$
$$\operatorname{cov}(M,Y) - \operatorname{cov}(T,M)\cdot\widetilde{\beta}_T - \operatorname{cov}(M,M)\cdot\beta_M = \beta_U\cdot\varphi_U + \beta_V\cdot\delta_Y \tag{74}$$
$$\operatorname{cov}(T,M) - \operatorname{cov}(T,T)\cdot\widetilde{\varphi}_T = \delta_T\cdot\xi_V \tag{75}$$

Simple manipulation of Equations (70)–(75) generate the identification of the following parameters:

$$\xi_Z = \frac{\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,Z)} \qquad \text{from Eq.(70)} \tag{76}$$
$$\widetilde{\varphi}_T = \frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)} \qquad \text{from Eq.(71)} \tag{77}$$
$$\beta_M = \frac{\operatorname{cov}(Z,T)\operatorname{cov}(T,Y) - \operatorname{cov}(T,T)\operatorname{cov}(Z,Y)}{\operatorname{cov}(T,M)\operatorname{cov}(Z,T) - \operatorname{cov}(T,T)\operatorname{cov}(Z,M)} \qquad \text{from Eqs.(72)–(73)} \tag{78}$$
$$\widetilde{\beta}_T = \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \qquad \text{from Eqs.(72)–(73)} \tag{79}$$
$$\beta_U\cdot\varphi_U + \beta_V\cdot\delta_Y = \operatorname{cov}(M,Y) - \operatorname{cov}(M,M)\cdot\beta_M - \operatorname{cov}(T,M)\cdot\widetilde{\beta}_T \qquad \text{from Eq.(74)} \tag{80}$$
$$\delta_T\cdot\xi_V = \frac{\operatorname{cov}(T,M)\operatorname{cov}(Z,M) - \operatorname{cov}(T,T)\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,M)} \qquad \text{from Eq.(75)} \tag{81}$$

Moreover, if we divide Equation (72) by $\operatorname{cov}(Z,T)$ we obtain:

$$\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} - \frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)}\cdot\beta_M - \frac{\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,T)}\cdot\widetilde{\beta}_T = 0 \tag{82}$$
$$\Rightarrow \quad \frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} - \widetilde{\varphi}_T\cdot\beta_M - \widetilde{\beta}_T = 0 \tag{83}$$
$$\Rightarrow \quad \widetilde{\varphi}_T\cdot\beta_M + \widetilde{\beta}_T = \frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}. \tag{84}$$

The four causal of interest parameters defined in (59)–(62) are respectively identified by Equations (77), (78), (79) and (84):

$$\frac{dE(M(t))}{dt} = \widetilde{\varphi}_T = \frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)}, \tag{85}$$
$$\frac{dE(Y(m))}{dm} = \beta_M = \frac{\operatorname{cov}(Z,Y)\operatorname{cov}(T,T) - \operatorname{cov}(Y,T)\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,M)\operatorname{cov}(T,T) - \operatorname{cov}(M,T)\operatorname{cov}(Z,T)}, \tag{86}$$
$$\frac{\partial E(Y(t,m))}{\partial t} = \widetilde{\beta}_T = \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)}, \tag{87}$$
$$\frac{dE(Y(t))}{dt} = \widetilde{\varphi}_T\cdot\beta_M + \widetilde{\beta}_T = \frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}. \tag{88}$$

Next section explains that each causal effect (85)–(88) can be evaluated by standard Two-stage Least Squares regressions.

Online Appendix E Exploring Alternative Approaches

We investigate the mediation model in which the treatment variable T and the mediator variable M are endogenous. Our solution imposes causal relations among unobserved variables that enable the identification of three causal effects using only one dedicated instrument for T.

Our method contrasts to two broad alternative approaches to gaining identification in mediation analysis. One of these is to assume that the treatment T and the mediator M are exogenous given observed variables (Imai et al., 2010, 2011a,b).[9] In this case, treatment T is as good as randomly assigned and the resulting model is equivalent to assuming no confounding variables and no unobserved mediators U in *Model III* of Table 1.[10] Relatedly, Yamamoto (2014) studies the case of a binary treatment indicator and a single instrument, assuming that the instrument Z is independent of the counterfactual outcome $Y(m,t)$ and that the mediator variables is exogenous conditioned on treatment compliance.[11]

A second class of models relies on additional instrumental variables dedicated to the mediator M. Burgess et al. (2015); Powdthavee (2009) and Jhun (2015) achieve identification using two instruments and parametric assumptions that shape the endogeneity of T and M. Two important contributions to this literature that use non-parametric identification are Frölich and Huber (2017) and Jun et al. (2016).[12] This second class of models does not assume away confounding effects; i.e. variables T, M, Y remain endogenous. It thus constitutes an alternative approach to our identification problem, which is to seek for another instrument that is dedicated to M.[13] Because of its natural appeal, we discuss this approach here and contrast its identification requirements explicitly to ours. A standard mediation model with confounding variables V and two separate dedicated instrumental variables (for separate endogenous variables) is described as follows:

$$\text{Treatment variable: } T = f_T(Z_T, V, \epsilon_T), \tag{89}$$

$$\text{Observed Mediator: } M = f_M(T, Z_M, V, \epsilon_M), \tag{90}$$

$$\text{Outcome: } Y = f_Y(T, M, V, \epsilon_Y), \tag{91}$$

$$\text{where: } (Z_T, Z_M) \perp\!\!\!\perp V. \tag{92}$$

This model is presented as a DAG in Table Online Appendix Table 2. In this model, the exclusion restriction $Z_M \perp\!\!\!\perp Y(m)$ and also $Z_M \perp\!\!\!\perp Y(m)|T$ hold. Thereby Z_M can be used to evaluate the causal effects of M on Y.[14]

The empirical challenge in evaluation Model (89)–(92) is to find a suitable candidate for Z_M. There are three potential concerns with any dedicated instrument for M: *(i)* Z_M may correlate with V; *(ii)* Z_M may directly affect Y; and *(iii)* Z_M may correlate with Z_T. Concerns *(i)* and *(ii)* define the usual requirements for any valid instrument to identify the effect of M on Y. The latter concern *(iii)* is specific to the mediation context. The three concerns are depicted as dashed errors in the right figure of Table Online Appendix Table 2. A potential candidate for Z_M is

[9] Robins and Greenland (1992) and Geneletti (2007) consider instruments that perfectly correlate with the mediator variable such that the exogeneity condition still holds.

[10] If the treatment T were indeed randomly assigned, then one could use the interaction of the treatment with observed covariates as instruments to identify the causal effect of M on Y. Versions of this approach are examined in Dunn and Bentall (2007); Gennetian et al. (2002); Ten Have et al. (2007); Small (2012).

[11] In our notation, this means that $Y(m,t) \perp\!\!\!\perp Z$ and $Y(t,m) \perp\!\!\!\perp M(t)|(T, P = c)$, where T denotes treatment assignment and P stands for an indicator of treatment compliance. Neither assumption holds in *Model III* or *Model IV* of Table 1.

[12] Both papers examine the effect of a binary indicator for treatment T. Frölich and Huber (2017) relies on two dedicated instruments (for T and M) and a monotonicity restriction with respect to M. Jun et al. (2016) uses three dedicated instruments but does not require the monotonicity restriction.

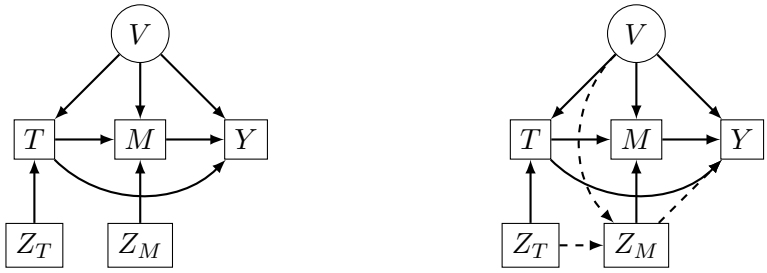
[13] Recently, Frölich and Huber (2017) provide an important contribution on the mediation model with two dedicated instruments.

[14] If T were to cause Z_M, then only $Z_M \perp\!\!\!\perp Y(m)|T$ would hold.

Table Online Appendix Table 2: General Mediation Model and Violation of Exclusion Restrictions

A. Directed Acyclic Graph (DAG) Representation

General IV Model with Two Instruments *Violations of the Exclusion Restriction*



The left figure gives the directed acyclic graph (DAG) representation of the general IV Model with two dedicated instruments. The right figure gives the same DAG, but also depicts the identification concerns discussed in the body of the text.

automation. Automation, i.e. replacing workers with machines, robots and computer-assisted technologies, is usually viewed as the 'other big shock' that has hit high-wage labor markets in the last decades. For example, Acemoglu and Restrepo (2017) estimate that an additional robot per thousand workers has reduced employment in the U.S. by about 0.18–0.34 percentage points and wages by 0.25–0.50 percent. The effects of automation are not expected to abate. For example, Frey and Osborne (2017) predict that 47% of U.S. workers are at risk of automation over the next two decades. In brief, automation has had and will likely continue to have substantial effects on labor market outcomes M and therefore seems like a good candidate dedicated instrument Z_M. We view concern *(iii)* as addressed in this context because Autor et al. (2015) have provided convincing evidence that automation and import exposure are largely orthogonal, making the two forces separable in the data at both the industry-level and the regional level. Concern *(i)* still is that firms may automate in response to other unobserved factors that could directly impact their labor demand. Indeed, firm-level technology upgrading does appear to respond to the China shock as shown by Bloom et al. (2016). This violates the independence $Z_M \perp\!\!\!\perp V$ in (92) and thereby the exclusion restriction $Z_M \perp\!\!\!\perp Y(m)|T$ does not hold. However, this concern may again be largely addressed if we think of Z_M not as actually measured technology upgrading but as some more exogenous measure, e.g. *exposure to robot adoption* as in Acemoglu and Restrepo (2017) or employment-weighted occupational measures like *routine task intensity* (Autor and Dorn, 2013) or *automatability* (Frey and Osborne, 2017). In our empirical context, concern *(ii)*—automation could impact voting behavior through channels other than M—is the most worrisome, and in fact clearly disqualifies automation as a dedicated instrument for M. While a German assembly-line worker will likely neither observe nor care about Australian imports of Chinese consumer electronics (i.e. Z_T), he/she will not only be aware of the potential automatability of their assembly-line job (i.e. Z_M) but may indeed seek out a more protectionist political agenda in anticipation of automation's consequences, i.e. even before any detrimental effects in the labor market.

Online Appendix F Estimation of Causal Parameters

Our goal is to show that the four causal parameters listed in Equations (85)–(88) can be estimated using the standard Two-stage Least Square (2SLS) estimator. We revise the standard equations of the 2SLS estimators for sake of completeness.

Equations (93)–(94) present the first and stages of a generic 2SLS regression in which T stands for the endogenous variable, Z is the instrumental variable and Y is the targeted outcome.

$$\text{First Stage: } T = \kappa_1 + \beta_1 \cdot Z + \epsilon_1, \tag{93}$$

$$\text{Second Stage: } Y = \kappa_2 + \beta_2 \cdot T + \epsilon_2. \tag{94}$$

The 2SLS estimator relies on the assumptions that the instrument Z is statistically independent of the term ϵ_2 while T is not. It is well-known that the 2SLS estimator $\hat{\beta}_2$ is given by the ratio of the sample covariances $\text{cov}(Z, Y)$ and $\text{cov}(Z, T)$. Moreover $\hat{\beta}_2$ is a consistent estimator of parameter β_2 :

$$\text{plim}(\hat{\beta}_2) = \frac{\text{cov}(Z, Y)}{\text{cov}(Z, T)} = \beta_2. \tag{95}$$

Consider the inclusion of additional covariates X in both stages of the 2SLS method. Variables X in (96)–(97) play the role of control covariates in the first stage and second stages of the 2SLS estimator. Control covariates X directly causes Y in (97) while the instrument Z only causes Y though it impact on T.

$$\text{First Stage: } T = \kappa_1 + \beta_1 \cdot Z + \psi_1 \cdot X + \epsilon_1, \tag{96}$$

$$\text{Second Stage: } Y = \kappa_2 + \beta_2 \cdot T + \psi_2 \cdot X + \epsilon_2. \tag{97}$$

The 2SLS model (96)–(97) relies on the assumption that the instrument Z and control covariates X are independent of error term ϵ_2, that is, $(Z, X) \perp\!\!\!\perp \epsilon_2$. The 2SLS estimator $\hat{\beta}_2$ for parameter β_2 is expressed by Equation (98) and it is a consistent estimator under model assumptions.

$$\text{plim}(\hat{\beta}_2) = \frac{\text{cov}(Z, Y)\,\text{cov}(X, X) - \text{cov}(Y, X)\,\text{cov}(Z, X)}{\text{cov}(Z, T)\,\text{cov}(X, X) - \text{cov}(T, X)\,\text{cov}(Z, X)} = \beta_2. \tag{98}$$

The 2SLS estimator $\hat{\psi}_2$ for parameter ψ_2 is expressed by Equation (99) and it is a consistent estimator under model assumptions.

$$\text{plim}(\hat{\psi}_2) = -\frac{\text{cov}(Z, Y)\,\text{cov}(T, X) - \text{cov}(Y, X)\,\text{cov}(Z, T)}{\text{cov}(Z, T)\,\text{cov}(X, X) - \text{cov}(T, X)\,\text{cov}(Z, X)} = \psi_2. \tag{99}$$

Each of the identification formulas for the causal effects in (85)–(88) describes a ratio of covariances that corresponds to one of the three 2SLS formulas (95), (98) or (98).

The effect of choice T on mediator M is given by:

$$\frac{dE(M(t))}{dt} = \widetilde{\varphi}_T = \frac{\text{cov}(Z, M)}{\text{cov}(Z, T)}.$$

According to Equation (95), this effect can be estimated by the 2SLS regression (93)–(94) in which Z is the instrument, T is the endogenous variable and M is the outcome.

The total effect of T on outcome Y is given by:

$$\frac{dE(Y(t))}{dt} = \widetilde{\varphi}_T \cdot \beta_M + \widetilde{\beta}_T = \frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}.$$

According to Equation (95),this effect can be estimated by the 2SLS regression (93)–(94) in which Z is the instrument, T is the endogenous variable and Y is the outcome.

The causal effect of mediator M on outcome Y is given by:

$$\frac{dE(Y(m))}{dm} = \beta_M = \frac{\operatorname{cov}(Z,Y)\operatorname{cov}(T,T) - \operatorname{cov}(Y,T)\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,M)\operatorname{cov}(T,T) - \operatorname{cov}(M,T)\operatorname{cov}(Z,T)},$$

which can be estimated by the 2SLS regression (93)–(94) where Z is the instrument, T is the endogenous variable and M is the outcome.

The causal effect of mediator M on outcome M is given by:

$$\frac{dE(Y(m))}{dm} = \beta_M = \frac{\operatorname{cov}(Z,Y)\operatorname{cov}(T,T) - \operatorname{cov}(Y,T)\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,M)\operatorname{cov}(T,T) - \operatorname{cov}(M,T)\operatorname{cov}(Z,T)}.$$

According to the 2SLS estimator in (98), this causal effect can be estimated by $\hat{\beta}_2$ in the 2SLS regression (96)–(97) in which Z plays the role of the instrument, M is the endogenous variable, T is the control covariate and Y is the outcome.

The Indirect Effect of choice T on outcome Y is given by:

$$\frac{\partial E(Y(t,m))}{\partial m} = \widetilde{\beta}_T = \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)}.$$

According to the 2SLS estimator in (99), this causal effect can be estimated by $\hat{\psi}_2$ in the 2SLS regression (96)–(97) in which Z plays the role of the instrument, M is the endogenous variable, T is the control covariate and Y is the outcome.

Online Appendix G Total, Indirect and Direct Effects under One Instrument

Online Appendix D.2 describes a linear mediation model whose primary causal effects are identified by the following equations:

$$\text{Total Effect of } T \text{ on } Y: \frac{dE(Y(t))}{dt} = \frac{\text{cov}(Z,Y)}{\text{cov}(Z,T)}. \tag{100}$$

$$\text{Direct Effect of } T \text{ on } Y: \frac{\partial E(Y(t,m))}{\partial t} = \frac{\text{cov}(Z,M)\,\text{cov}(T,Y) - \text{cov}(Z,Y)\,\text{cov}(T,M)}{\text{cov}(T,T)\,\text{cov}(Z,M) - \text{cov}(Z,T)\,\text{cov}(T,M)}. \tag{101}$$

$$\text{Effect of } M \text{ on } Y: \frac{\partial E(Y(t,m))}{\partial m} = \frac{\text{cov}(Z,T)\,\text{cov}(T,Y) - \text{cov}(T,T)\,\text{cov}(Z,Y)}{\text{cov}(T,M)\,\text{cov}(Z,T) - \text{cov}(T,T)\,\text{cov}(Z,M)} \tag{102}$$

$$\text{Effect of } T \text{ on } M: \frac{dE(M(t))}{dt} = \frac{\text{cov}(Z,M)}{\text{cov}(Z,T)} \tag{103}$$

$$\text{Indirect Effect of } T \text{ on } Y: \frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt}. \tag{104}$$

The literature of mediation analysis typically expresses the total effect of T on Y as the sum of its direct and indirect effects. In our notation, this decomposition is is stated as following:

$$\underbrace{\frac{dE(Y(t))}{dt}}_{\text{Total Effect}} = \underbrace{\frac{\partial E(Y(t,m))}{\partial t}}_{\text{Direct Effect}} + \underbrace{\frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt}}_{\text{Indirect Effect}}. \tag{105}$$

We show that the decomposition described in (105) is exact in the case of a single instrument. That is to say that the covariance ratio that identifies the total effect of T on Y in equation (100) is equal to the covariance ratio that identifies the direct effect in Equations (101) plus the multiplication of the covariance ratios that identify the effect of T on M in (103) and the effect of

M on Y described in Equation (102). We thank David Slichter for pointing out this fact.

$$
\begin{aligned}
&\underbrace{\frac{\partial E(Y(t,m))}{\partial t}}_{\text{Direct Effect}} + \underbrace{\frac{\partial E(Y(t,m))}{\partial m} \cdot \frac{dE(M(t))}{dt}}_{\text{Indirect Effect}} \\
&= \underbrace{\frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)}}_{\frac{\partial E(Y(t,m))}{\partial t}} + \underbrace{\frac{\operatorname{cov}(Z,T)\operatorname{cov}(T,Y) - \operatorname{cov}(T,T)\operatorname{cov}(Z,Y)}{\operatorname{cov}(T,M)\operatorname{cov}(Z,T) - \operatorname{cov}(T,T)\operatorname{cov}(Z,M)}}_{\frac{\partial E(Y(t,m))}{\partial m}} \cdot \underbrace{\frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)}}_{\frac{dE(M(t))}{dt}} \\
&= \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} + \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(T,T)\operatorname{cov}(Z,Y)\frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)}}{\operatorname{cov}(T,M)\operatorname{cov}(Z,T) - \operatorname{cov}(T,T)\operatorname{cov}(Z,M)} \\
&= \frac{\operatorname{cov}(Z,M)\operatorname{cov}(T,Y) - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} + \frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,Y)\frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)} - \operatorname{cov}(Z,M)\operatorname{cov}(T,Y)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \\
&= \frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,Y)\frac{\operatorname{cov}(Z,M)}{\operatorname{cov}(Z,T)} - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \\
&= \frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,M)\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \\
&= \frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,M)\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} - \operatorname{cov}(Z,Y)\operatorname{cov}(T,M)\frac{\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,T)}}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \\
&= \frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,M)\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)} \\
&= \left(\frac{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)}{\operatorname{cov}(T,T)\operatorname{cov}(Z,M) - \operatorname{cov}(Z,T)\operatorname{cov}(T,M)}\right) \cdot \left(\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}\right) \\
&= \frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)} = \underbrace{\frac{dE(Y(t))}{dt}}_{\text{Total Effect}}.
\end{aligned}
$$

The first equality expresses the total effect of T on Y in terms of its direct and indirect effects. The second equality substitutes the direct and indirect effects by their identification formulas described in (101), (102) and (100). The third equation isolates and eliminates the common term $\operatorname{cov}(Z,M)$ in the denominator of $\frac{dE(Y(m))}{dm}$. The fourth equation flips the sign of the terms in the last covariance ratio. Now the overall sum has the same denominator. The fifth equation eliminates the common term in the sum of the numerators of both ratios. The sixth equation exchange the covariances $\operatorname{cov}(Z,M)$ and $\operatorname{cov}(Z,Y)$ of the first term of the numerator. The seventh equation includes the term $\frac{\operatorname{cov}(Z,T)}{\operatorname{cov}(Z,T)}$ which is equal to one. The eight equation exchange the covariances $\operatorname{cov}(Z,Y)$ and $\operatorname{cov}(Z,T)$ of the second term of the numerator. The night equation isolates the common denominator of the expression. The tenth equation eliminates the common first term of both numerator and denominator. The resulting formula is the covariate ratio $\frac{\operatorname{cov}(Z,Y)}{\operatorname{cov}(Z,T)}$ which, according to (100), is equal to the total effect of choice T on outcome Y.

References

Acemoglu, Daron and Pascual Restrepo, "Robots and Jobs: Evidence from US Labor Markets.," *MIT Umpublished Mimeo.*, 2017.

Autor, David and David Dorn, "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market," *American Economic Review*, 2013, *103* (5), 1553–1597.

_ , _ , and Gordon H Hanson, "Untangling trade and technology: Evidence from local labour markets," *The Economic Journal*, 2015, *125* (584), 621–646.

Bloom, Nicholas, Mirko Draca, and John Van Reenen, "Trade induced technical change? The impact of Chinese imports on innovation, IT and productivity," *The Review of Economic Studies*, 2016, *83* (1), 87–117.

Burgess, S., R. M. Daniel, A. S. Butterworth, and S. G. Thompson, "Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways," *International Journal of Epidemiology*, 2015, *44* (2), 484–495.

Dunn, G. and R. Bentall, "Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments)," *Statistics in Medicine*, 2007, *26* (26), 4719–4745.

Frey, Carl B. and Michael A. Osborne, "The Future of Employment: How Susceptible are Jobs to Computerisation?," *Oxford Martin School Umpublished Mimeo.*, 2017.

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, pp. n/a–n/a.

Geneletti, S., "Identifying direct and indirect effects in a non-counterfactual framework," *Journal of the Royal Statistical Society B*, 2007, *69* (2), 199–215.

Gennetian, L. A., J. Bos, and P. Morris, "Using Instrumental Variables Analysis to Learn More from Social Policy Experiments," MDRC Working Papers on Research Methodology, MDRC (Manpower Demonstration Research Corporation) 2002.

Have, T. R. Ten, M. M. Joffe, K. G. Lynch, G. K. Brown, S. A. Maisto, and A. T. Beck, "Causal mediation analyses with rank preserving models," *Biometrics*, September 2007, *63* (3), 926–934.

Heckman, James J., "The Principles Underlying Evaluation Estimators with an Application to Matching," *Annales d'Economie et de Statistiques*, 2008, *91–92*, 9–73.

Heckman, James J and Rodrigo Pinto, "Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs," *Econometric reviews*, 2015, *34* (1-2), 6–31.

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

_ , _ , Dustin Tingley, and Teppei Yamamoto, "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies," *American Political Science Review*, 2011, *105*, 765–789.

_ , _ , _ , **and** _ , "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies," *American Political Science Review*, 2011, *105* (4), 765–789.

Jhun, M. A., "Epidemiologic approaches to understanding mechanisms of cardiovascular diseases: genes, environment, and DNA methylation." PhD dissertation, University of Michigan, Ann Arbor 2015.

Jun, Sung Jae, Joris Pinkse, Haiqing Xu, and Neşe Yildiz, "Multiple Discrete Endogenous Variables in Weakly-Separable Triangular Models," *Econometrics*, 2016, *4* (1).

Pearl, Judea, "The Mediation Formula: A Guide to the Assessment of Causal Pathways in Nonlinear Models," 2011. Forthcoming in *Causality: Statistical Perspectives and Applications*.

Petersen, M. L., S. E. Sinisi, and M. J. Van der Laan, "Estimation of direct causal effects," *Epidemiology*, 2006, *17*, 276–284.

Powdthavee, Nattavudh, "Does Education Reduce Blood Pressure? Estimating the Biomarker Effect of Compulsory Schooling in England," Discussion Paper 09/14, University of York, Department of Economics, York, UK 2009.

Robins, J. M., "Semantics of causal DAG models and the identification of direct and indirect effects.," in N. L. P. J. Green, Hjort and S. Richardson, eds., *Highly Structured Stochastic Systems*, MR2082403, Oxford: Oxford University Press, 2003, pp. 70–81.

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

Rosenbaum, Paul R. and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, April 1983, *70* (1), 41–55.

Rubin, D. B., "Direct and indirect causal effects via potential outcomes (with discussion)," *Scandinavian Journal of Statistics*, 2004, *31*, 161–170.

Small, D. 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.

Yamamoto, T., "Identification and estimation of causal mediation effects with treatment noncompliance," March 2014. Manuscript. Department of Political Science, Massachusetts Institute of Technology, Cambridge.