Mediation Analysis in IV Settings With a Single Instrument*

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

July 14, 2020

Abstract

Economists have long used instrumental variables to evaluate the causal effect of an endogenous treatment variable on outcomes of interest. Most empirical settings consist of a limited number of instrumental variables for a relatively large number of outcomes. The standard IV model enables us to employ a single instrumental variable to identify the causal effects of a treatment variable on multiple outcomes. The IV model, however, is not suitable to identify the causal effect of an intermediate outcome on a final outcome. Such a task is often called mediation analysis. We address the question of whether it is possible to use the same instrumental variable to identify the causal chain amongst outcomes in a standard IV model without revoking the endogeneity of the treatment with respect to intermediate and final outcomes. We show that mediation effects can be identified when the IV model is partially confounded, i.e. the unobserved confounding variables that cause the treatment and the intermediate outcome are independent of the confounders that cause the intermediate and final outcomes. This assumption generates additional exogeneity conditions without altering the key features of the IV model. We discuss the intuition, plausibility, and estimation of the partially confounded IV model. Finally, we illustrate a range of practical problems faced by empirical economists to show where the partially confounding condition holds – and where it does not.

Keywords: Instrumental Variables, Mediation Analysis, Causal Effects, Identification.
JEL Codes: C36

*A version of this paper was presented at the conference celebrating the contribution of James Heckman to Economics, held at University of Chicago in June, 2019. We thank Edward Vytlacil, Alex Torgovitzky, Steven Durlauf, James Heckman, Sonia Bhalotra, Johanna Fajardo, Andreas Ferrara, Markus Frölich, Martin Huber, Kosuke Imai, Ed Leamer, Yi Lu, Craig McIntosh, Bruno Pellegrino, Giacomo Ponzetto, 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 primary task of policy evaluation is to assess the causal effect of an intervention, i.e. a treatment, on outcomes of interest. A fundamental problem for the identification of causal effects with observational data is endogeneity, when unobserved confounding variables cause both the treatment variable and the outcome. Endogeneity induces a correlation between the treatment variable and the outcome which prevents the identification of causal effects. A popular solution to address the problem of endogenous treatment is the use of instrumental variables (IV) . In this paper, we explore whether the standard IV model can be extended beyond the treatment effects. We investigate if it is possible to identify the causal relations among different outcomes while relying on a single instrument.

IV models are characterised by an exclusion restriction stating that an instrument is an exogenous variable that affects the outcome only through its impact on the treatment variable. The exclusion restriction implies that the IV is independent of counterfactual outcomes. This statistical relationship is called the exogeneity condition and can be employed to identify the causal effect of a treatment on various outcomes.[1]

The exclusion restriction limits the empirical availability of good IVs. A typical empirical IV setting consists of a treatment of interest T, a single or a limited number of instruments Z, and a relatively large number of outcomes Y. The same instrument Z can be used to identifying the effect of T on multiple Y. A natural extension of this analysis is to evaluate the causal effects among outcomes. Specifically, the researcher may seek to evaluate the causal chain in which a treatment T causes an intermediate outcome M which in turn causes a final outcome Y. Identifying this causal chain implies disentangling the total causal effect of treatment T on final outcome Y into a direct effect of T on Y and the indirect effects operating through one or several intermediate outcomes M. This task is often called mediation analysis and it has numerous applications in the literature of applied economics. For instance, Heckman et al. (2013) evaluate how a policy intervention during childhood (T) affects early cognitive and non-cognitive skills (M), and how this affects labor market outcomes in adulthood (Y). Similarly, Flores and Flores-Lagunes (2010) evaluate how a training intervention on labor market skills affects the pursue of higher schooling degrees which in turn,

[1] Note that the exogeneity condition alone is not sufficient for the identification of causal effects. An extensive IV literature exists on the additional assumptions that render the identification of causal effects.

enhance earnings at later ages.

Despite its benefits, mediation analysis is rare in IV evaluations. The difficulty in assessing causal mechanisms is due to the fact that the standard IV model is not suitable to investigate the causal effects among outcomes. Specifically, the exogeneity condition that allows identifying causal treatment effects on both intermediate and final outcomes does not identify the causal relation between these outcomes.

This paper addresses the question whether it is possible to perform mediation analysis with a single set of instrumental variables while preserving the essential properties of the IV model. A typical IV model enables to employ Z to separately identify the causal effect of a treatment T on an intermediate outcome M, and the effect of T on a final outcome Y. We investigate conditions that enable us to use the same instrument Z to assess the causal effect of intermediate outcome M on final outcome Y, while maintaining the endogeneity of the treatment T with respect to the intermediate outcome M and final outcome Y. We show that the nonparametric identification of mediation effects is possible if the IV model is *partially confounded*. This means that unobserved confounding variables that jointly cause the treatment and the intermediate outcome are independent of the confounders that cause the intermediate and the final outcome.

We discuss the interpretation and limitations of the partially confounded IV model. We exemplify cases where the model holds and discuss empirical settings that violate its assumptions. We present estimation procedures for two popular empirical models. The first refers to nonparametric identification and estimation of mediation effects for the case of a binary treatment and a binary mediator. We extend the Local Instrumental Variable Model (LIV) of Heckman and Vytlacil (1999) to perform mediation analysis. The second considers a linear model as it is often used in applications, and employs a single IV in several Two Stage Least Squares (2SLS) regressions to identify mediation effects.

This paper adds to the literature on the identification of causal effects using instrumental variables. It offers a practical guide for the empirical economist that employs IVs and seeks to go beyond treatment effects. If the empirical setting is such that the partially confounding condition is justified, then the researcher can use our method for assessing causal effects among different outcomes of the same treatment variable. If the condition does not hold, then the economist may pursue additional instrumental variables for the intermediate outcome or opt to evaluate bounds

for the mediation effects.

This paper also contributes to the literature on mediation analysis that employs instrumental variables. Robins and Greenland (1992), Mattei and Mealli (2011), Imai et al. (2013) and Attanasio et al. (2020) study the case of an instrument for the intermediate outcome under the assumption of an exogenous treatment. Dunn and Bentall (2007), Albert (2008), Small (2012), Chen et al. (2019), Joffe et al. (2008) investigate the case of an instrument for the intermediate outcome and an endogenous treatment. They invoke parametric assumptions such as linearity to identify mediation effects. Frölich and Huber (2017) investigate the case of two instrumental variables, one for the treatment and another one for the intermediate outcome. Joffe et al. (2008) assume a single instrument that jointly affects the treatment and the intermediate outcome. Similar to our setting, Brunello, Fort, and Winter-Ebmer (2016) and Yamamoto (2014) consider the case of a single instrumental variable for an endogenous binary treatment. Brunello et al. (2016) evoke an exogeneity assumption on observed covariates to control for the endogeneity of the intermediate outcome. Yamamoto (2014) controls for endogeneity of the intermediate outcome by assuming that the mediator is exogenous when conditioned on treatment compliance. His approach can be understood as assuming that the confounding variables causing T, M are conditionally independent of the confounding variables causing T, Y. We contribute to this literature by investigating the assumptions that enable the identification of mediation effects without altering the statistical features that characterize the standard IV model. We employ a single IV for the treatment while maintaining the endogenous property of the treatment with respect to the mediator and the final outcome.

The rest of the paper proceeds as follows. Section 2 describes the mediation model with IV and discusses its identification challenges. Section 3 examines exogeneity conditions generated by relaxing the dependence structure among error terms that can be used for addressing this challenge. Specifically, Section 3.2 examines the statistical properties of the mediation model with partially confounded error terms, which allows for identification. Section 4 discusses the interpretation of the partially confounded IV model with mediators. Section 5 extends to the LIV model of Heckman and Vytlacil (1999) to evaluate mediation effects of the partially confounded model for a binary treatment and a binary mediator. Section 6 examines the linear IV model and shows that mediation effects can be identified by standard 2SLS regressions. Section 7 describes extensions of the basic

partially confounded model that maintain its exogeneity conditions. Section 8 concludes.

2 The IV Model with an Intermediate Outcome as Mediator

The goal of mediation analysis is to disentangle the total effect (TE) of treatment T on outcome Y into two components. The indirect effect (IE) is the effect of T on Y that operates exclusively through the impact that T has on an intermediate outcome called the mediator M.[2] The direct effect (DE) is the causal effect that the treatment T would have on Y if the distribution of the mediator M were held constant. Consider a binary treatment where T takes values in $\text{supp}(T) = \{t_0, t_1\}$, and let $Y(t, m)$ be the counterfactual outcome when the treatment T is fixed at a value $t \in \{t_0, t_1\}$ and M is fixed at a value $m \in \text{supp}(M)$. The total effect (TE), direct effect (DE) and indirect effect (IE) are defined[3] by:

$$TE = E(Y(t_1) - Y(t_0)) \equiv \int E(Y(t_1, M(t_1)) - Y(t_0, M(t_0))), \tag{1}$$

$$DE(t) = E(Y(t_1, M(t)) - Y(t_0, M(t))) \equiv \int E\Big(Y(t_1, m) - Y(t_0, m)\Big) dF_{M(t)}(m), \tag{2}$$

$$IE(t) = E(Y(t, M(t_1)) - Y(t, M(t_0))) \equiv \int E\Big(Y(t, m)\Big)\Big[dF_{M(t_1)}(m) - dF_{M(t_0)}(m)\Big], \tag{3}$$

where $F_{M(t)}(m) = P(M(t) \leq m)$ stands for the cumulative distribution of counterfactual mediation variable $M(t); t \in \{t_0, t_1\}$.[4]

The total effect stands for the average causal effect of the treatment T on outcome Y including its impact on the mediator M. The direct effect evaluates the share of the total effect that does *not* operate via M. It contemplates the expected outcome difference that would occur due to an exogenous change in T while setting M to its counterfactual values $M(t)$. The indirect effect evaluates the expected change in the outcome when the treatment is fixed at value t while the mediator is exogenously manipulated from $M(t_0)$ to $M(t_1)$. The total effect can be decomposed as

[2] The applied literature frequently refers to the IE as effect mechanism, i.e. the observed channel through which a broader treatment effect actually affects Y.

[3] Pearl (2012, 2014) make a distinction between controlled (or "prescriptive") and natural (or "descriptive") effects. He uses the terms *controlled* direct effect and *controlled* indirect effect for equations (2) and (3) respectively.

[4] These equations are termed the mediation formula.

the sum of the direct and indirect effect (Robins and Greenland, 1992):

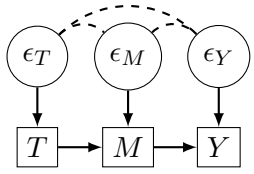
$$TE = DE(t_1) + IE(t_0) \text{ or } TE = DE(t_0) + IE(t_1). \tag{4}$$

Pearl (2001) introduced the term *controlled direct effect* (CDE) for the average effect of T on Y when M is fixed at a value $m \in \text{supp}(M)$:

$$CDE(m) = E\Big(Y(t_1, m) - Y(t_0, m)\Big). \tag{5}$$

We investigate the identification of mediation effects using a general mediation model as shown in Table 1. The first column describes the model equations where error terms $\epsilon_T, \epsilon_M, \epsilon_Y$ consist of unobserved random vectors. The second column displays the Directed Acyclic Graph (DAG) associated with the model.[5] The last column displays the counterfactual (potential) mediation $M(t)$ for T fixed at $t \in \text{supp}(T)$ and the counterfactual outcome $Y(t, m)$ for (T, M) fixed at values $(t, m) \in \text{supp}(T) \times \text{supp}(M)$. Throughout this paper, we suppress additional covariates X for sake of notational simplicity. All analyses can be understood as being conditional on X.

Table 1: The Mediation Model

Model Equations	*DAG*	*Counterfactual Variables*
$T = f_T(\epsilon_T)$ $M = f_M(T, \epsilon_M)$ $Y = f_Y(T, M, \epsilon_Y)$ $\epsilon_T \not\perp\!\!\!\perp \epsilon_M,\ \epsilon_T \not\perp\!\!\!\perp \epsilon_Y,\ \epsilon_M \not\perp\!\!\!\perp \epsilon_Y$		$M(t) = f_M(t, \epsilon_M)$ $Y(m, t) = f_Y(t, m, \epsilon_Y)$

Mediation effects could be easily identified if error terms $\epsilon_T, \epsilon_M, \epsilon_Y$ were mutually independent. In this case, it easy to show that $T, M(t), Y(m, t)$ are also mutually independent:

$$T \perp\!\!\!\perp M(t), \quad T \perp\!\!\!\perp Y(m, t), \text{ and } M(t) \perp\!\!\!\perp Y(m, t). \tag{6}$$

[5] Throughout this paper, we will use DAGs for graphical representation of econometric models. Causal relations are depicted by arrows, dashed lines denote statistical dependency, circles stand for unobserved variables while squares denote observed variables. It is worth noting that DAGs only serve a didactic purpose. They do not add any additional information into the model and are only used as a visualization tool.

The independence relationship $Y(t,m) \perp\!\!\!\perp (T, M(t))$ suffices to identify $E(Y(t,m))$:

$$E(Y(t,m)) = E(Y(t,m)|T=t, M(t)=m) = E(Y|T=t, M=m),$$

where the first equality is due to $Y(t,m) \perp\!\!\!\perp (T, M(t))$ and the last one comes from the fact that $Y(T,M) = Y$ and $M(T) = M$. The identification of $E(Y(t,m))$ is not sufficient to obtain the direct and indirect effects. These effects require the identification of the expectation $E(Y(t, M(t')))$. Under (6), $Y(t,m) \perp\!\!\!\perp (M(t), T)$ holds and $E(Y(t, M(t')))$ can be identified by:

$$\begin{aligned}
E(Y(t,M(t'))) &= \int E\big(Y(t,m)|M(t')=m\big)dF_{M(t')}(m) && \text{by Law of Iterated Expectations} \\
&= \int E\big(Y(t,m)|M(t)=m, T=t\big)dF_{M(t')}(m) && \text{by } Y(t',m) \perp\!\!\!\perp (M(t),T) \\
&= \int E\big(Y(t,m)|T=t, M(t)=m\big)dF_{M|T=t'}(m) && \text{by } M(t') \perp\!\!\!\perp T \\
&= \int E\big(Y|T=t, M=m\big)dF_{M|T=t'}(m) && \text{from } Y \equiv Y(T,M(T)) \text{ and } M \equiv M(T)
\end{aligned}$$

A large mediation literature relies on the exogeneity assumptions as in (6) to identify mediation effects. Imai et al. (2010), for example, use the term Sequential Ignorability for the following independence assumptions:[6]

$$\big(Y(t',m), M(t)\big) \perp\!\!\!\perp T|X \tag{7}$$

$$Y(t',m) \perp\!\!\!\perp M(t)|(T,X) \tag{8}$$

Assumptions (7)–(8) are equivalent to (6) as they imply $(T, M(t')) \perp\!\!\!\perp Y(t,m)$ when baseline variables X are suppressed.[7]

The independence assumptions in (6) are rather strong. They state that no unobserved variable can jointly cause T, M, Y and thereby, they rule out any possibility of endogeneous effects between these variables. A natural quest of the mediation literature is to weaken this assumption. One

[6]The Sequential Ignorability assumption of Imai et al. (2010) is stated as conditioned on pre-treatment variable X. As mentioned before, we usually suppress X for sake of notational simplicity.

[7]According to the Graphoid Axiom of Contraction (Lauritzen, 1996) we have that:

$$Y(t',m) \perp\!\!\!\perp T \text{ and } Y(t',m) \perp\!\!\!\perp M(t)|T \Rightarrow Y(t',m) \perp\!\!\!\perp (M(t),T).$$

Therefore we have that $Y(t',m) \perp\!\!\!\perp M(t)$ and $Y(t',m) \perp\!\!\!\perp T$ hold. Moreover, (7) implies that $M(t) \perp\!\!\!\perp T$ and thereby $T, M(t)$ and $Y(t,m)$ are mutually independent.

possibility is to employ two instrumental variables to the mediation model in Table 1, one that causes T and another one that causes M. The instrument that causes T can be used to address the endogeneity of treatment T with respect to M, Y, while the instrument that causes M can address the endogeneity of M w.r.t. Y. This approach is investigated in Frölich and Huber (2017). The requirement of two instrumental variables poses an empirical burden as it is often difficult to find a valid instrument for T alone.

The focus of this paper differs slightly from the main inquiry of the mediation literature that employs instrumental variables. Our primary object of analysis is not the mediation model described above but the standard IV model. We are not seeking to investigate how additional instruments in the mediation model enable identification. Instead, we examine how additional assumptions in the standard IV model identify mediation effects while maintaining model endogeneity. The scope of this search is constrained to the IV model properties, namely, we employ a single instrument for T and the identifying assumptions cannot rule out the endogeneity concerns that called for an instrument in the first place.

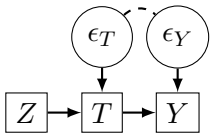
2.1 Standard IV Model

The standard IV model does not include mediators. It stems from three observed variables: an instrumental variable Z, that causes a treatment variable T, which in turn causes an outcome Y. Table 2 summarizes the key features of the standard IV model. The first column displays the model equations. The causal relations between Z, T and Y are determined by the unknown functions $f_T(\cdot)$ and $f_Y(\cdot)$. Unobserved error terms ϵ_T, ϵ_Y consist of random vectors that are arbitrarily correlated. They play the role of confounding variables that cause both the treatment T and outcome Y. Error terms are assumed to be jointly independent of the exogenous instrumental variable Z, that is, $Z \perp\!\!\!\perp (\epsilon_T, \epsilon_Y)$.

The second column of Table 2 displays the IV model as a DAG. The third column presents the counterfactual treatment choice $T(z)$ when Z is fixed at a value $z \in \text{supp}(Z)$ and the counterfactual outcome $Y(t)$ for T fixed at a value $t \in \text{supp}(T)$. The last column presents the independence relations of the IV model. The exogeneity condition states that $Z \perp\!\!\!\perp (T(z), Y(t))$, which is a consequence of $Z \perp\!\!\!\perp (\epsilon_T, \epsilon_Y)$. The statistical dependence between error terms renders T endogenous w.r.t. outcome Y. This means that T is not statistically independent of counterfactual outcomes

$Y(t)$. Notationally, we use $T \not\perp\!\!\!\perp Y(t)$ to indicate that T is endogenous w.r.t. outcome Y.

Table 2: Standard IV Model

Model Equations	*DAG*	*Counterfactuals*	*Model Properties*
$T = f_T(Z, \epsilon_T)$ $Y = f_Y(T, \epsilon_Y)$ $Z \perp\!\!\!\perp (\epsilon_T, \epsilon_Y)$		$T(z) = f_T(z, \epsilon_T)$ $Y(t) = f_Y(t, \epsilon_Y)$	$Z \perp\!\!\!\perp T(z)$ $Z \perp\!\!\!\perp Y(t)$ $T \not\perp\!\!\!\perp Y(t)$

The exogeneity condition $Z \perp\!\!\!\perp (T(z), Y(t))$ of the IV model is necessary but not sufficient to identify the causal effect of T on Y. The identification of treatment effects requires additional assumptions that may vary across a range of possibilities. We employ some of these techniques to identify mediation effects of the partially confounded IV model that is derived in further sections. A widely used identification assumption is linearity, which enables the evaluation of treatment effects by Two-Stage Least Squares (2SLS) regressions. An undesirable feature of the linearity assumption is that the treatment effect is homogeneous across the individuals of the targeted population.

An extensive literature in econometrics offers weaker assumption that allow for treatment heterogeneity. Heckman and Vytlacil (2005) investigate the case of a binary treatment assignment and assume that the treatment assignment is characterized by a threshold-crossing function that is separable in Z and ϵ_T.[8] Imbens and Angrist (1994) invoke a monotonicity criterion on counterfactual choices that enables the identification of the the Local Average Treatment Effect.[9] Vytlacil (2002) shows that monotonicity and separability assumptions are equivalent. Heckman and Pinto (2017) present an unordered monotonicity condition that applies to unordered choice models with multiple treatments. Pinto (2015) investigates identifying assumptions generated by revealed preference analysis. Lee and Salanié (2015) investigate the identification of treatment effects in categorical choice models under an arbitrary set of threshold-crossing rules. Altonji and Matzkin (2005); Blundell and Powell (2003, 2004); Imbens and Newey (2007); Matzkin (2003) offer identification results that employ the method of control functions.[10]

[8] Namely $T = \mathbf{1}[\phi(Z) \geq \xi(\epsilon_T)]$, where $\mathbf{1}[\cdot]$ denotes a indicator function.

[9] They assume that for any $z, z' \in \text{supp}(Z)$, $T_i(z) \geq T_i(z')$ for all agents $i \in \mathcal{I}$ or $T_i(z) \geq T_i(z')$ for all agents $i \in mathcalI$.

[10] For instance, Imbens and Newey (2007) show that for T continuously distributed, if the function $f_T(Z, \epsilon_T)$ is strictly increasing in ϵ_T, then the conditional CDF $V(z,t) = P(T \leq t | Z = z)$ can be used as a balancing score for ϵ_T, which yields the exogenous condition $Y(t) \perp\!\!\!\perp T | V(Z, T)$.

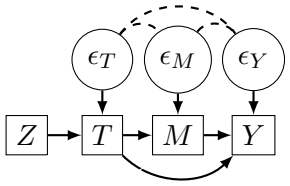
In summary, there are several possibilities to exploit the properties of the IV model in order to identify treatment effects. The seminal feature of the IV model shared by these methods is that T is endogenous w.r.t. Y and that Z is exogenous w.r.t. $T(z), Y(t)$. We refer to these two conditions using the following notation:

$$\text{IV Model Properties:} \quad \underbrace{T \not\perp\!\!\!\perp Y(t)}_{T \text{ endogenous w.r.t. } Y} \quad \text{and} \quad \underbrace{Z \perp\!\!\!\perp (T(z), Y(t))}_{Z \text{ is exogenous w.r.t. } Y(t), T(z)} . \tag{9}$$

2.2 The Mediation Model with IV

We now expand the standard IV model by adding a mediator M that plays the role of an intermediate outcome caused by T which is known to cause the final outcome Y. Table 3 describes a mediation model with IV in its most general form. The first column displays the model equations which determine the causal relations among variables. Error terms $\epsilon_T, \epsilon_M, \epsilon_Y$ denote unobserved random vectors with an arbitrary dependence relation. Error terms are assumed to be statistically independent of the exogenous instrumental variable Z. The second column displays the model as a DAG.

Table 3: IV Model with Mediator Variable

Model Equations	*DAG*
$T = f_T(Z, \epsilon_T)$ $M = f_M(T, \epsilon_M)$ $Y = f_Y(T, M, \epsilon_Y)$ $Z \perp\!\!\!\perp (\epsilon_T, \epsilon_M, \epsilon_Y)$	

There are six counterfactual variables that stem from the mediation model in Table 3. These variables are listed in the first column of Table 4. Rows 1–2 display counterfactual choice T and mediator M when Z is fixed at some value $z \in \text{supp}(Z)$. Rows 3–4 display counterfactual mediator M and final outcome Y when T is fixed. Row 5 displays the counterfactual outcome when only the mediator M is fixed while row 6 displays the counterfactual outcome Y when the mediator M and choice T are fixed at $(m, t) \in \text{supp}(M) \times \text{supp}(T)$.

The second column of Table 4 indicates whether the counterfactual variable is statistically inde-

pendent of the instrumental variable Z. These relations are a direct consequence of the independence between instrument Z and the error terms, that is $Z \perp\!\!\!\perp (\epsilon_T, \epsilon_M, \epsilon_Y)$. The last row states that the instrumental variable Z and the counterfactual outcome $Y(m)$ are not statistically independent. Indeed, $Y(m)$ is a function of T which is caused by Z. The last column of Table 4 examines if outcomes are endogenous. It shows that no counterfactual outcome (M or Y) is statistically independent of the variables it is fixed upon. These relations stem from the arbitrary dependence structure among error terms $\epsilon_T, \epsilon_M, \epsilon_Y$.

Table 4: Counterfactual Variables of the Primary Mediator Model

	Counterfactuals	*IV Relation*	*Outcome Endogeneity*
1.	$T(z) = f_T(z, \epsilon_T)$	$T(z) \perp\!\!\!\perp Z$	–
2.	$M(z) = f_M(T(z), \epsilon_M)$	$M(z) \perp\!\!\!\perp Z$	–
3.	$M(t) = f_M(t, \epsilon_M)$	$M(t) \perp\!\!\!\perp Z$	$M(t) \not\perp\!\!\!\perp T$
4.	$Y(t) = f_Y(t, M(t), \epsilon_Y)$	$Y(t) \perp\!\!\!\perp Z$	$Y(t) \not\perp\!\!\!\perp T$
5.	$Y(m) = f_Y(T, m, \epsilon_Y)$	$Y(m) \not\perp\!\!\!\perp Z$	$Y(m) \not\perp\!\!\!\perp M$
6.	$Y(t, m) = f_Y(t, m, \epsilon_Y)$	$Y(t, m) \perp\!\!\!\perp Z$	$Y(t, m) \not\perp\!\!\!\perp (T, M)$

This table describes a range of counterfactual variables in the primary mediation model of Table 3. Each counterfactual variable can be expressed in terms of the instrument Z and erros terms $\epsilon_T, \epsilon_M, \epsilon_Y$ by iterated substitution. The independence relations stem from $Z \perp\!\!\!\perp (\epsilon_T, \epsilon_M, \epsilon_Y)$ while the lack of statistical independence comes from the fact that error terms share an arbitrary dependence relation.

Examining the Causal Effects of T on M and T on Y

The mediation model of Table 3 embeds two standard IV models. The sub-model comprising the observed variables Z, T, M and error terms ϵ_T, ϵ_M complies with the standard IV model of Table 2. Rows 1 and 3 of Table 4 display the exogenous and endogenous conditions that characterize an standard IV model for the mediator variable, namely:

$$\underbrace{T \not\perp\!\!\!\perp M(t)}_{T \text{ endogenous w.r.t. } M} \quad \text{and} \quad \underbrace{Z \perp\!\!\!\perp (T(z), M(t))}_{\text{Exogeneity Condition for } M(t)} \tag{10}$$

The sub-model of variables Z, T, Y can be obtained by suppressing the mediator M using iterated substitutions. As expected, this sub-model also constitutes a standard IV model. Rows 1 and 4 of

Table 4 generate its exogeneity and endogeneity conditions:

$$\underbrace{T \not\perp\!\!\!\perp Y(t)}_{T \text{ endogenous w.r.t. } Y} \quad \text{and} \quad \underbrace{Z \perp\!\!\!\perp (T(z), Y(t))}_{\text{Exogeneity Condition for } Y(t)} \tag{11}$$

Condition (10) enable the use of the instrumental variable Z to identify the effect of T on M while condition (11) is useful in the identification of the total effect of T on Y.

Examining the Joint Effect of T, M on Y

The last row of Table 4 states that T, M are endogenous with respect to Y (third column). Rows 1,2 and 6 of Table 4 imply the exogeneity condition $Z \perp\!\!\!\perp (Y(z,t), M(z), T(z))$, which means that the instrument Z can be used to evaluate the joint effect of (M, T) on Y. To clarify, consider a variable $G = f_G(T, M)$ where $f_G : \text{supp}(T) \times \text{supp}(M) \to \mathcal{G}$ that is a one-to-one mapping from the values that T, M take into an indexing set $\mathcal{G}$. We can rewrite the outcome $Y = f_Y(T, M, \epsilon_Y)$ as $Y = g_Y(G, \epsilon_Y)$ without loss of generality. Let the counterfactual treatment be $G(z) = f_G(T(z), M(z))$ and the counterfactual outcome be $Y(g) = g_Y(g, \epsilon_Y)$. This model complies with the IV exogeneity property that $Z \perp\!\!\!\perp G(z), Y(g)$ and the instrument Z could be used to evaluate the joint effect of T, M on Y via G. Although we could evaluate the joint effect of T, M on Y, we cannot disentangle these effects. In particular, we cannot distinguish the direct of T on Y from the indirect effect that operates through M. The difficulty of decomposing these effects hinges on the problem of identifying the causal effect of M on Y.

Examining the Causal Effect of M on Y

The fifth row of Table 4 states that $Y(m) \not\perp\!\!\!\perp M$, which means that the Mediator M is endogenous w.r.t. Y. A natural inquiry is to examine if the instrumental variable Z can be employed to identify this effect. Unfortunately, the exogeneity condition $Y(m) \perp\!\!\!\perp Z$ does not hold because $Y(m)$ is a function of T which is caused by Z. In summary, we have that:

$$\underbrace{M \not\perp\!\!\!\perp Y(m)}_{M \text{ endogenous w.r.t. } Y} \quad \text{and} \quad \underbrace{Z \perp\!\!\!\perp M(z), \text{ but } Z \not\perp\!\!\!\perp Y(m)}_{Z \text{ is not exogenous w.r.t. } Y(m)}. \tag{12}$$

We can also investigate if Z can be used to evaluate the causal effect of M on Y when condi-

tioning on T.[11] Unfortunately, the general error dependence does not render Z a valid instrument for identifying the effect of M on Y either. Indeed, conditioning on $T = t$ means that the values of Z and ϵ_T are such that $f_T(Z, \epsilon_T) = t$ holds. This induces a statistical dependency between Z and ϵ_T. Moreover, the dependence between ϵ_T and ϵ_Y implies that Z and ϵ_Y are not statistically independent when conditioning on $T = t$. But $Y(m)$ is a function of ϵ_Y and thereby we have that:

$$Y(m) \not\perp\!\!\!\perp Z|T \tag{13}$$

In summary, the general dependence relation among error terms $\epsilon_T, \epsilon_M, \epsilon_Y$, invalidates Z as an instrumental variable for the causal effect of M on Y regardless of whether we condition on T or not. Intuitively, it means that changes in Z do not yield an exogenous variation in M that could be used to identify its effect on Y. The next section discusses this identification challenge in more detail.

2.3 The Challenge of Disentangling the Causal Effect of T, M on Y

The challenge of isolating the causal effect of M on Y can be understood as the difficulty of generating exogenous variation in M while controlling for confounding variables. We clarify this challenge by referring to the control function method.[12] Consider the following assumption:

Assumption A-1. ϵ_T, ϵ_M are continuous random variables with strictly increasing cumulative density functions (CDFs) $F_{\epsilon_T}(e) = P(\epsilon_T \leq e), F_{\epsilon_M}(e) = P(\epsilon_M \leq e)$.

Let the CDF transformations of the error terms ϵ_T, ϵ_M be $U_T = F_{\epsilon_T}(\epsilon_T)$ and $U_M = F_{\epsilon_M}(\epsilon_M)$. Under Assumption **A-1**, U_T and U_M are uniformly distributed in $[0, 1]$, that is $U_T \sim unif[0, 1]$, $U_M \sim unif[0, 1]$. Moreover, U_T, U_M constitute a one-to-one mapping of their respective error terms. We can now restate the choice and mediation equations of Table 4 in terms of U_T and U_M :

$$T = f_T(Z, U_T) \tag{14}$$

$$M = f_M(T, U_M) \tag{15}$$

[11] Conditioning on T will be a relevant feature of the partially confounded IV model we are going to derive later.

[12] See Matzkin (2003) and Imbens and Newey (2009) for a discussion of identification results using the control function approach.

The model is completed by the outcome equation $Y = f_Y(T, M, \epsilon_Y)$ and the independence condition $Z \perp\!\!\!\perp (U_T, U_M, \epsilon_Y)$. We seek to identify a set of control variables C that render the mediator M independent of counterfactual outcome $Y(m)$, that is $Y(m,t) \perp\!\!\!\perp (M,T)|C$ holds. To do so, we rely on strong functional form assumptions:

Assumption A-2. Let T be a scalar treatment variable and $f_T(z,u)$ be a strictly increasing function in u.

Let $F_{T|Z}(t,z) = P(T \leq t|Z = z)$ be the propensity score CDF and let Z_i, T_i, M_i, Y_i be the observed variables of an individual i that are treated as non-stochastic values. Under Assumption **A-2**, U_T can be identified as $U_T = F_{T|Z}(T,Z)$ as follows:

$$\begin{aligned}
F_{T|Z}(T_i, Z_i) &= P(T \leq T_i|Z = Z_i) \\
&= P(f_T(Z_i, U_T) \leq f_T(Z_i, U_{T,i})|Z = Z_i)) \\
&= P(f_T(U_T \leq U_{T,i}|Z = Z_i) \quad \text{due to } \textbf{A-2} \\
&= P(f_T(U_T \leq U_{T,i}) \quad \text{due to } Z \perp\!\!\!\perp U_T \\
&= U_{T,i} \quad \text{due to } U_T \sim Unif[0,1]
\end{aligned}$$

Variable U_T is a control variable for the causal effect of T on M and Y. That is to say that $M(t) \perp\!\!\!\perp T|U_T$ and $Y(t) \perp\!\!\!\perp T|U_T$ hold.[13] Under full support of the propensity score, we can identify the mean of the counterfactual mediator $E(M(t))$ and counterfactual outcome $E(Y(t))$ by:

$$E(M(t)) = \int_0^1 E(M|T = t, U_T = u)du \quad \text{and} \quad E(Y(t)) = \int_0^1 E(Y|T = t, U_T = u)du.$$

Intuitively, identification arises because we can use the instrument Z to vary U_T while keeping $T = t$ unchanged. We explore the same rationale to investigate a control variable for error term U_M.

Assumption A-3. Let M be a scalar mediator, and $f_M(t,u)$ be a strictly increasing function in u.

[13] The independence $(U_T, U_M, \epsilon_Y) \perp\!\!\!\perp Z$ implies that $(U_M, \epsilon_Y) \perp\!\!\!\perp Z|U_T$. Choice $T = f_T(U_T, Z)$, is a function of Z when conditioned on U_T and the counterfactual $M(t) = f_M(t, U_M)$ is a function of U_M. Thus $U_M \perp\!\!\!\perp Z|U_T$ implies that $M(t) \perp\!\!\!\perp T|U_T$. In the same fashion, $Y(t) = f_Y(t, M(t), \epsilon_Y) = f_Y(t, f_M(t, U_M), \epsilon_Y)$ is a function of U_M, ϵ_Y and therefore $(U_M, \epsilon_Y) \perp\!\!\!\perp Z|U_T$ implies that $Y(t) \perp\!\!\!\perp T|U_T$.

Assumption A-4. Let the conditional CDF $F_{U_M|U_T}(u_m, u_T) = P(U_M \leq u_m | U_T = u_t)$ be strictly increasing in u_m.

Let the CDF of M conditioned on Z, U_T be $F_{M|Z,U_T}(m, z, u) = P(M \leq m | Z = z, U_T = u)$. Consider the random variable $C = F_{M|Z,U_T}(M, Z, U_T)$. We claim that (U, C) is a one-to-one mapping of (U_T, U_M). Let $C_i = F_{M|Z,U_T}(M_i, Z_i, U_{T,i})$ be the value of variable C for individual i. Thus we have that:

$$\begin{aligned}
C_i = F_{M|Z,U_T}(M_i, Z_i, U_{T,i}) &= P(M \leq M_i | Z = Z_i, U_T = U_{T,i}) \\
&= P(f_M(f_T(Z_i, U_{T,i}), U_M) \leq f_M(f_T(Z_i, U_{T,i}), U_{M_i}) | Z = Z_i, U_T = U_{T,i})) \\
&= P(U_M \leq U_{M_i} | Z = Z_i, U_T = U_{T,i})) \quad \text{due to } \textbf{A-3} \\
&= P(U_M \leq U_{M_i} | U_T = U_{T,i})) \quad \text{due to } Z \perp\!\!\!\perp U_M | U_T
\end{aligned}$$

Assumption **A-4** assures that (U_T, C) is a one-to-one transformation of (U_T, U_M). For instance, consider the values $(U_{T,i}, U_{M,i})$ and $(U_{T,j}, U_{M,j})$ are such that $U_{T,i} = U_{T,j}$ and $U_{M,i} < U_{M,j}$. **A-4** implies that $P(U_M \leq U_{M_i} | U_T = U_{T,i})) < P(U_M \leq U_{M_j} | U_T = U_{T,j}))$ and thereby $C_i < C_j$. As a consequence, conditioning on (U_T, C) is equivalent to conditioning on (U_T, U_M).

We employ control variables (U_T, C) to investigate the identification of the counterfactual outcome $Y(t, m)$. Variables T, M are a function of Z, U_T, U_M, and conditioning on U_T, U_M renders M, T as a function only of Z. On the other hand, $Y(t, m) = f_Y(t, m, \epsilon_Y)$ is a function of ϵ_Y. The independence relationship $Z \perp\!\!\!\perp \epsilon_Y | (U_T, U_M)$ implies that $Y(t, m) \perp\!\!\!\perp (T, M) | (U_T, U_M)$, which can be equivalently stated as $Y(t, m) \perp\!\!\!\perp (T, M) | (U_T, C)$. Thus control variables (U_T, C) enable to identify the conditional expectation of the counterfactual outcome by $E(Y(t, m) | U_T = u, C = c) = E(Y | T = t, M = m, U_T = u, C = c)$.

The outcome mean could be obtained by integrating $E(Y | T = t, M = m, U_T = u, C = c)$ over the distribution of (U_T, C) across its support, that is, $E(Y(t, m)) = \int_0^1 \int_0^1 E(Y | T = t, M = m, U_T = u, C = c) dF_{U_T,C}(u, c)$. Unfortunately, the model does not allow to generate the exogenous variation on both U_T and C that is necessary to evaluate the integral of Y when conditioning on M and T. The choice equation relates to three variables T, Z and U_T. Given a value $T = t$, we can vary Z to scan the values that U takes in its support. However, C is a function of M, Z, U and it is

deterministic when conditioned on $M = m$ given the values of Z and U_T. In other words, we have no degrees of freedom to manipulate the values of the control variable C.

It is worth noting that **A-2** and **A-3** are strong assumptions. They require T and M to be continuous variables. Our goal here is not to defend the model assumptions, but to gain intuition on the challenge of simultaneously identifying all causal effects in the mediation model when only one instrumental variable Z is available. We will revisit the control functions approach later in the paper to exemplify the application of an additional assumption that grants identification.

3 Examining Independence Assumptions on the Error Terms

The previous section has investigated the general mediation model that allows for an arbitrary dependence between error terms $\epsilon_T, \epsilon_M, \epsilon_Y$. The section has shows that the effect of M on Y cannot be identified without additional assumptions. We now examine if relaxing the independence conditions among error terms enables the identification of the effect of the mediator on the outcome while maintaining the endogeneity of the treatment variable. Equations (16)–(19) describe the basic features of our mediation model.

$$T = f_T(Z, \epsilon_T) \tag{16}$$
$$M = f_M(T, \epsilon_M) \tag{17}$$
$$Y = f_Y(T, M, \epsilon_Y) \tag{18}$$
$$Z \perp\!\!\!\perp (\epsilon_T, \epsilon_M, \epsilon_Y) \tag{19}$$

According to Table 4, the instrumental variable is statistically independent of the following counterfactual variables: $Z \perp\!\!\!\perp (T(z), M(z), M(t), Y(t), Y(t, m))$. Mediator M and the counterfactual outcome $Y(m)$ are a function of T and thereby $M \not\perp\!\!\!\perp Y(m)$ always holds. Instrumental variable Z causes treatment T that in turn causes the counterfactual outcome $Y(m)$, thereby $Z \not\perp\!\!\!\perp Y(m)$ always holds.

Our task is three-fold: we seek to (1) relax the degree of association among error terms $\epsilon_T, \epsilon_M, \epsilon_Y$; (2) generate exogeneity conditions that help to identify the causal effect of M on Y; (3) while maintaining T endogenous.

By relaxing the dependance among error terms we mean that we investigate the assumptions $\epsilon_T \perp\!\!\!\perp \epsilon_M$, $\epsilon_T \perp\!\!\!\perp \epsilon_Y$, or $\epsilon_M \perp\!\!\!\perp \epsilon_Y$. The exogeneity conditions that would aid in the identification

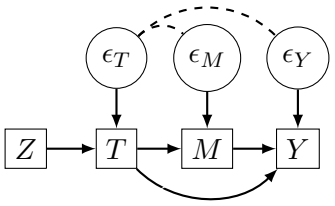
of the effect of M on Y comprise a matching condition $M \perp\!\!\!\perp Y(m)|T$, the advent of a control variable C such that $M \perp\!\!\!\perp Y(m)|(T, C)$ holds, or an IV exogeneity condition that renders Z a valid instrument to access the causal effect of M on Y, that is, $Z \perp\!\!\!\perp Y(m)|T$.[14] Maintaining the endogeneity of the treatment variable means that $T \not\perp\!\!\!\perp M(t)$ and $T \not\perp\!\!\!\perp Y(t)$ must hold.

Assuming that $\epsilon_T \perp\!\!\!\perp \epsilon_M$, renders $M(t) = f_M(t, \epsilon_M)$ and $T = f_T(Z, \epsilon_T)$ statistically independent, which violates the endogeneity of T with respect to M. Therefore our task summarizes to examine the properties of the two mediation models generated by assuming $\epsilon_M \perp\!\!\!\perp \epsilon_Y$ or $\epsilon_T \perp\!\!\!\perp \epsilon_Y$.[15]

3.1 Investigating the assumption of independence between ϵ_M and ϵ_Y

We examine the properties of the IV-mediation model (16)–(19) under the assumption that $\epsilon_M \perp\!\!\!\perp \epsilon_Y$. We show the independence assumption does not violate the endogeneity of the treatment T w.r.t. Y or M, but does not render Z a valid instrument for the causal effect of M on Y either. The model is presented as a DAG in Table 5.

Table 5: IV Model under the Assumption that $\epsilon_M \perp\!\!\!\perp \epsilon_Y$

DAG	*Key Properties*
	$T \not\perp\!\!\!\perp (M(t), Y(t))$ $M \not\perp\!\!\!\perp Y(m)\|T$ $Z \not\perp\!\!\!\perp Y(m)\|T$

Endogenous Treatment

The assumption $\epsilon_M \perp\!\!\!\perp \epsilon_Y$ does not imply independence between error terms ϵ_M, ϵ_Y and ϵ_T, which can share an arbitrary dependence relation. Notionally, we write, $\epsilon_T \not\perp\!\!\!\perp (\epsilon_M, \epsilon_Y)$. The treatment choice T is a function of ϵ_T while $M(t), Y(t)$ are a function of ϵ_M, ϵ_Y respectively, therefore, $T \not\perp\!\!\!\perp (M(t), Y(t))$. This means that the treatment choice T remains endogenous under the assumption $\epsilon_M \perp\!\!\!\perp \epsilon_Y$.

Seeking Exogeneity Conditions

[14] According to the model features, $Z \perp\!\!\!\perp Y(m)$ never holds.

[15] Assuming that $\epsilon_M \perp\!\!\!\perp \epsilon_Y$ **and** $\epsilon_T \perp\!\!\!\perp \epsilon_Y$ does not generate a palatable model as it implies that $Y(t, m) \perp\!\!\!\perp (T, M)$ and therefore the causal effect of T, M on Y can be identified by simply conditioning on these variables, namely, $E(Y(m, t)) = E(Y|M = m, T = t)$.

The assumption $\epsilon_M \perp\!\!\!\perp \epsilon_Y$ does not imply the exogeneity condition $Z \perp\!\!\!\perp Y(m)|T$. Conditioning on $T = t$ implies conditioning on the values of Z, ϵ_T such that $f_T(Z, \epsilon_T) = t$ holds, which induces a dependence relation between Z and ϵ_T. Error term ϵ_T shares a dependence structure with ϵ_M, ϵ_Y which implies that $Z \not\perp\!\!\!\perp (\epsilon_T, \epsilon_M, \epsilon_Y)|T$. In particular, we have that $Z \not\perp\!\!\!\perp Y(m)|T$ as $Y(m)$ is a function of ϵ_Y. Consequently, $\epsilon_M \perp\!\!\!\perp \epsilon_Y$ does not render Z a valid instrument for the effect of M on Y.

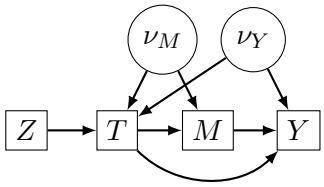
The assumption $\epsilon_M \perp\!\!\!\perp \epsilon_Y$ does not imply the matching condition $M \perp\!\!\!\perp Y(m)|T$ either. Although ϵ_M ϵ_Y are statistically independent, these error terms are not independent conditional on ϵ_T, that is, $\epsilon_M \not\perp\!\!\!\perp \epsilon_Y|\epsilon_T$, and thereby we also have that $\epsilon_M \not\perp\!\!\!\perp \epsilon_Y|T$, which implies that $M \not\perp\!\!\!\perp Y(m)|T$.

Confounding Variables and a Balancing Score for T

It is useful to relate assumption $\epsilon_M \perp\!\!\!\perp \epsilon_Y$ to some relevant literature on both IV and mediation analysis. To do so, we restate the error terms $\epsilon_T, \epsilon_M, \epsilon_Y$ as a function of unobserved random vectors ν_M, ν_Y such that $\epsilon_M = \xi_M(\nu_M)$, $\epsilon_Y = \xi_Y(\nu_Y)$, $\epsilon_T = \xi_T(\nu_M, \nu_Y)$. In this notation, the independence assumption $\epsilon_M \perp\!\!\!\perp \epsilon_Y$ translates into $\nu_M \perp\!\!\!\perp \nu_Y$ and we can express T as a function of ν_M, ν_Y, that is $T = f_T(Z, \nu_T, \nu_Y)$. Table 6 restates our model in terms of the new error terms. The DAG in Table 6 shows that the independence assumption can be interpreted as stating that the confounding variables that jointly affect T, Y are independent of the confounding variables that jointly cause T, M.

Table 6: Restating the IV Model Using Joint Confounding Variables

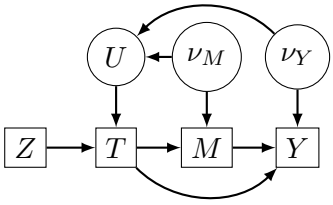
$T = f_T(Z, \nu_M, \nu_Y)$
$M = f_M(T, \nu_M)$
$Y = f_Y(T, M, \nu_Y)$
Z, ν_M, ν_Y are mutually independent



Consider a balancing score $U = \phi(\nu_M, \nu_Y)$ such that $T = f_T(Z, U)$. Balancing score U reduces the dimension of error term ν_M, ν_Y in the choice equation and relates to several identification strategies. It represents a control function in the case of a continuous treatment discussed in Section 2.3. In the Local Instrumental Variable (LIV) model of Heckman and Vytlacil (2001), the treatment choice indicator is given by $T = \mathbf{1}[P(Z) \geq U]$ where $P(Z)$ stands for the propensity score

and U is an unobserved variable that is uniformity distributed in $[0,1]$. In the LATE model of Imbens and Angrist (1994) where Z takes values in $\{z_0, z_1\}$, U stands for the vector of counterfactual choices $U = [T(z_0), T(z_1)]$ which describes the response-types of never-takers, always-takers, compliers and definers.[16] Table 7 displays the model equations, its associated DAG and some independence relations conditional on U.

Table 7: Restating the IV Model Using Joint Confounding Variables and Variable U

Model Equations	*DAG*	*Statistical Properties*
$U = f_U(\nu_M, \nu_Y)$ $T = f_T(Z, U)$ $M = f_M(T, \nu_M)$ $Y = f_Y(T, M, \nu_Y)$ Z, ν_M, ν_Y are mutually independent		$T \perp\!\!\!\perp (M(t), Y(t))\|U$ $T \perp\!\!\!\perp (Y(t,m), M(t))\|U$ $Y(t,m) \not\perp\!\!\!\perp M(t')\|(T,U)$

Conditional on U, treatment T only depends on Z, which is independent of ν_M, ν_Y. Therefore, $T \perp\!\!\!\perp (M(t), Y(t))|U$ holds and the causal effect of T on M, Y can be identified by conditioning on U.

Yamamoto (2013) identifies mediation effects invoking an assumption that combines the monotonicity condition of the LATE model in Imbens and Angrist (1994) with the sequential ignorability of Imai et al. (2010). We now investigate whether the independence assumption between ν_M, ν_Y justifies his assumptions. Yamamoto (2013) investigates the case of a binary treatment $T \in \{t_0, t_1\}$ and a binary instrument $Z \in \{z_0, z_1\}$. In our notation, his assumption can be stated as:

$$T \perp\!\!\!\perp (Y(t,m), M(t'))|U, \tag{20}$$

$$Y(t,m) \perp\!\!\!\perp M(t')|(T,U), \tag{21}$$

where $U = [T(z_0), T(z_1)]'$ stand for the response-types. Under monotonicity, the response-types are never-takes ($U = [t_0, t_0]'$), complies ($U = [t_0, t_1]'$) or always-takes ($U = [t_1, t_1]'$).

Equations (20)–(21) states that the sequential ignorability assumption (7)–(8) holds when conditioning on U. The independence assumption $\nu_M \perp\!\!\!\perp \nu_Y$ is not required for (20) to hold. The assumption simply exploits the control function property of U. Equation (21) states that the coun-

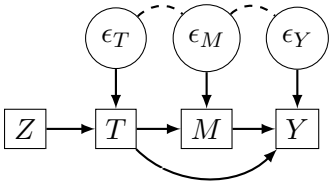
[16] In the general case of a multiple choice model with categorical instrument, U stands for the response-types, namely, the vector of counterfactual choices as Z ranges in its support.

terfactual outcome $Y(t,m)$ and the counterfactual mediator $M(t')$ are independent conditional on T and U. This assumption is not justified by the error terms' independence $\nu_T \perp\!\!\!\perp \nu_Y$. Note that $Y(t,m)$ is a function of ν_Y and $M(t')$ is a function of ν_M. Thus $\nu_Y \perp\!\!\!\perp \nu_M$ implies that $Y(t,m) \perp\!\!\!\perp M(t')$. However, ν_Y and ν_M are not independent conditional on T or U, and thereby $Y(t,m) \not\perp\!\!\!\perp M(t')|(T,U)$.

3.2 Investigating the Assumption of independence between ϵ_T and ϵ_Y

We now examine the properties of the IV-mediation model (16)–(19) under the assumption that $\epsilon_T \perp\!\!\!\perp \epsilon_Y$. The independence assumption does not violate the endogeneity of the treatment T w.r.t. Y or M. Moreover, it does render Z a valid instrument for the causal effect of M on Y. It turns out that $\epsilon_T \perp\!\!\!\perp \epsilon_Y$ is the only assumption on the error terms that comply to all the properties we desire. The assumption generates an exogeneity condition that enables the identification the effect of M on Y with the same Z used for identifying T on Y as well as T on M, without giving up the assumption on the endogeneity of T. Table 8 lists the primary statistical properties of this model.

Table 8: IV Model under the Assumption that $\epsilon_T \perp\!\!\!\perp \epsilon_Y$

DAG	*Statistical Properties*
	$T \not\perp\!\!\!\perp (M(t), Y(t))$ $M \not\perp\!\!\!\perp Y(m)\|T$ $Z \perp\!\!\!\perp Y(m)\|T$

Endogenous Treatment

The statistical dependence ϵ_T and ϵ_M implies that $T \not\perp\!\!\!\perp M(t)$. The counterfactual outcome is given by $Y(t) = f_Y(t, M(t), \epsilon_Y)$ which is a function of $M(t)$ and therefore $T \not\perp\!\!\!\perp Y(t)$ as well. Thus the treatment T remains endogenous under the assumption $\epsilon_T \perp\!\!\!\perp \epsilon_Y$.

The statistical dependence between ϵ_M and ϵ_Y implies that $M \not\perp\!\!\!\perp Y(m)|T$. This means that M is endogenous w.r.t. Y when conditioning on T. Therefore the causal effect of M on Y cannot be identified in the absence of an instrument Z. Conditioning on M does not render T exogenous with respect to its effect on Y either. Conditioning on M induces a correlation between error terms

ϵ_T and ϵ_Y, which implies that $T \not\perp\!\!\!\perp Y(t)|M$. However conditioning on T generates an exogenous condition that is useful to identify the effect of M on Y, as we show next.

New Exogenous Condition

Under assumption $\epsilon_T \perp\!\!\!\perp \epsilon_Y$, the variables $\epsilon_T, \epsilon_Y, Z$ become mutually independent. This implies that $Z \perp\!\!\!\perp \epsilon_Y|(f(Z,\epsilon_T) = t)$ holds, and thereby $Z \perp\!\!\!\perp f_Y(t, m, \epsilon_Y)|(f(Z,\epsilon_T) = t)$ also holds. As a consequence, we obtain a new exogenous condition $Z \perp\!\!\!\perp Y(m)|T$, which enables us to use Z as an instrument for M when conditioning on T. Assumption $\epsilon_T \perp\!\!\!\perp \epsilon_Y$ enables to use a single IV Z (that directly causes T) to aid in the identification of three causal effects, $T \to M$, $T \to Y$, and $M \to Y$.

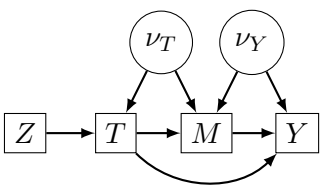
Although $Z \perp\!\!\!\perp Y(m)|T$ holds, $Z \perp\!\!\!\perp Y(m)$ does not. Conditioning on T is a necessary condition to use Z as instrument to evaluate the causal effect of M on Y. Intuitively, $Z \perp\!\!\!\perp Y(m)|T$ means that conditional on T, a change in the instrument Z generates an exogenous variation of the mediator M that can be used to identify the causal effect of M on Y. It is useful to reframe the model to clarify this property.

Understanding the New Exogenous Condition

We can gain model intuition by expressing the error terms $\epsilon_T, \epsilon_M, \epsilon_Y$ as functions of two unobserved random vectors ν_T, ν_Y such that $\epsilon_T = \xi_T(\nu_T)$. $\epsilon_M = \xi_M(\nu_T, \nu_Y)$, $\epsilon_Y = \xi_Y(\nu_Y)$. The independence assumption $\epsilon_T \perp\!\!\!\perp \epsilon_Y$ translates to $\nu_T \perp\!\!\!\perp \nu_Y$ and M becomes a function of ν_T, ν_Y.

Table 9 presents the mediation model using error terms ν_T, ν_Y. According to its DAG, the independence assumption $\nu_T \perp\!\!\!\perp \nu_Y$ can be interpreted as stating that the confounding variable ν_T that jointly affects T, M is independent of the confounding variable ν_Y that jointly causes M, Y. We therefore term this model as *partially confounded*.

Table 9: The Partially Confounded IV Model

Model Equations	*DAG*	*Statistical Properties*
$T = f_T(Z, \nu_T)$ $M = f_M(T, \nu_T, \nu_Y)$ $Y = f_Y(T, M, \nu_Y)$ $Z \perp\!\!\!\perp (\nu_T, \nu_Y)$ and $\nu_T \perp\!\!\!\perp \nu_Y$		$T \not\perp\!\!\!\perp (M(t), Y(t))$ $M \not\perp\!\!\!\perp Y(m)\|T$ $Z \perp\!\!\!\perp Y(m)\|T$

The DAG in Table 9 is useful to interpret the exogeneity condition $Z \perp\!\!\!\perp Y(m)|T$. Consider the sub-model that suppresses variables Z and T. The resulting model can be understood as an IV model where the effect of M on Y is confounded by error term ν_Y, and the error term ν_T plays the role of an instrumental variable. If ν_T was observed, it could be used as an IV for the causal relation between M and Y. The error term ν_T is unobserved and it is unconditionally independent of the observed instrument Z. However, conditioning on T generates a statistical dependence between Z and ν_T. Hence variation in Z induces a variation in ν_T, which plays the role of a new instrumental variable for the causal effect of M on Y.

A stylized example provides some intuition on how Z affects ν_T when conditioned on T. Let T be a binary treatment that indicates college enrollment. Let M be a binary indicator that takes value $M = 1$ if the student belongs to an athletic or sport association on campus and $M = 0$ otherwise. The researcher is interested in identifying how much of the total effect of college T on adulthood income Y is mediated by attending the athletic/sport club M.[17]

Suppose students commute to college by walking and let the instrumental variable be the distance between their home and campus: $Z = 0$ indicates a short distance while $Z = 1$ indicates a long distance. Let the confounding variable ν_T indicate athletic ability, $\nu_T = 1$ for athletic students and $\nu_T = 0$ for sedentary ones. College distance Z and athletic ability ν_T are statistically independent. However, conditional on going to college, students that live far from college ($Z = 1$) are more likely to be athletic ($\nu_T = 1$). In other words, conditioning on college enrollment T generates a positive correlation between college distance Z and athletic ability ν_T.

Control Function Exercise Using the Partial Confounding Assumption

We resort back to the control function approach in Section 2.3 to facilitate the understanding of the partially confounded model of Table 9, We invoke assumptions that match those of Section 2.3:

Assumption A-1′. ν_T, ν_M are continuous random variables with strictly increasing cumulative density functions (CDFs) $F_{\nu_T}(e) = P(\nu_T \leq e), F_{\nu_Y}(e) = P(\nu_Y \leq e)$.

Let U_T, U_Y be the CDF transformation of error terms ν_T, ν_Y and the model equations are given by $T = f_T(Z, U_T)$, $M = f_M(T, U_T, U_Y)$, and $Y = f_Y(T, M, U_Y, \epsilon_Y)$. The exogeneity of the IV is

[17] The social contacts made at the sports club could help in finding a better job after college.

expressed by $Z \perp\!\!\!\perp (U_T, U_Y, \epsilon_Y)$ and the partial confounding assumption is given by $U_T \perp\!\!\!\perp (U_Y, \epsilon_Y)$. We insert the error term ϵ_Y to avoid a degenerated distribution of outcome Y. Error terms U_Y, ϵ_Y share an arbitrary dependence structure.

Assumption A-2′. Let T be a scalar treatment variable and $f_T(z, u)$ be a strictly increasing function in u.

Similar to Section 2.3, Assumption **A-2′** enables us to identify $U_T = F_{T|Z}(T, Z)$. Counterfactual means $E(M(t)) = \int_0^1 E(M|T = t, U_T = u)du$ and $E(Y(t)) = \int_0^1 E(Y|T = t, U_T = u)du$ are identified by using the variation of the instrument Z to assess the values of U_T while conditioning on $T = t$.

The analysis differs from the one in Section 2.3 as the partial confounding assumption enables us to identify U_Y. The identification employs two statistical relationships listed below: A useful property of the model is $(Z, U_T, T) \perp\!\!\!\perp (U_Y \epsilon_Y)$. Indeed, IV independence and the partial confounding assumption imply that $(Z, U_T) \perp\!\!\!\perp (U_Y \epsilon_Y)$ and the fact that $T = f(Z, U_T)$ implies $(Z, U_T, T) \perp\!\!\!\perp (U_Y \epsilon_Y)$. Now consider the assumption:

Assumption A-3′. Let M be a scalar mediator and $f_M(t, u, v)$ be a strictly increasing function in v.

Let the CDF of M conditioned on T, Z be $F_{M|T,Z}(m, t, z) = P(M \leq m|T = t, Z = z)$. Let $U_{T,i} = F_{T|Z}(T_i, Z_i)$ and $C_i = F_{M|T,Z}(M_i, T_i, Z_i)$ be the value of variable C for individual i. Thus:

$$
\begin{aligned}
C_i = F_{M|T,Z}(M_i, Z_i, T_i) &= P(M \leq M_i|T = t_i, Z = Z_i) \\
&= P(M \leq M_i|T = t_i, Z = Z_i, U_T = U_{T,i}) \\
&= P(f_M(T_i, T, U_{T,i}, U_Y) \leq f_M(f_T(Z_i, U_{T,i}), U_{Y_i})|T = t_i, Z = Z_i, U_T = U_{T,i})) \\
&= P(U_Y \leq U_{Y_i})) = U_{Y_i} \quad \text{due to } (Z, U_T, T) \perp\!\!\!\perp U_Y
\end{aligned}
$$

Under this partial confounding assumption, U_Y is identified by $U_Y = F_{M|T,Z}(M, Z, T)$. We can use the independence $(U_Y, \epsilon_Y) \perp\!\!\!\perp (Z, U_T, T)$ to generate another useful property, namely, $\epsilon_Y \perp\!\!\!\perp (Z, T, U_T)|U_Y$ holds and therefore we have that $Y(m, t) \perp\!\!\!\perp (M, T)|U_Y$, as $Y(m, t)$ is a function of ϵ_Y

and (M, T) are a function of (Z, T, U_T). In particular, we have that $Y(m) \perp\!\!\!\perp M|(T, U_Y)$ also holds. Thus we can identify the counterfactual mean $E(Y(m,t)) = E(Y(m)|T = t) = \int_0^1 E(Y|M = m, T = t, U_Y = u)du$. Empirically, we are able to identify $E(Y(m,t))$ because we can use the instrument Z to screen the values of U_Y when conditioned on $M = m$ and $T = t$.

The intuition for identification relies on the fact that the partial confounding assumption $U_T \perp\!\!\!\perp U_Y$ and the exogeneity condition $Y(m) \perp\!\!\!\perp Z|T$ are intertwined. On the one hand, the independence $U_T \perp\!\!\!\perp U_Y$ enables us to achieve identification by conditioning only on U_Y. On the other hand, $Y(m) \perp\!\!\!\perp Z|T$ enables us to exploit the variation of the instrumental variable Z to evaluate U_Y.

4 When does Partial Confounding Apply?

The partially confounded IV model of Table 9 is not always appropriate to describe the empirical data. The assumption $\nu_T \perp\!\!\!\perp \nu_Y$ is violated whenever the observed variables T, M, Y are jointly caused by the same unobserved confounding variable. How likely this is depends on the concrete empirical setting. For instance, consider evaluating the effect of a student's choice of majors on her income that is mediated by college grades. Let T be the choice of college majors, M be the grades of the students during college and let Y be the labor market income after college. We can make the case that unobserved cognitive ability is a common confounding factor that jointly causes the choice of majors, grades, and income, which violates the independence assumption between error terms.

The partially confounded model is better suited for empirical applications where the causal relations $T \to M$ and $T \to Y$ result from different processes (e.g. political versus economic activities), or relate to different agents (e.g. businesses versus individuals). We illustrate this fact using a couple of examples.

Suppose a policy maker intends to evaluate the impact of extra-lessons on college enrollment in poor school districts. She then devises a program that sends additional teachers to randomly chosen high-schools to give extra-lessons. The policy maker is interested in evaluating the impact of the program on college enrolment that operates via improved schooling grades. Apart from this indirect effect, the extra-lessons also advertise the benefits of college education, which has a direct effect on college enrollment by increasing the number of applications. In this setup, the randomization plays

the role of a binary instrument Z. Treatment T stands for the content or number of extra-lessons. The mediator M consists of school grades and the outcome Y is college enrollment. Error term ϵ_T captures unobserved skills of teachers who decided to join the program. Grades could improve due to better teacher quality or due to the content and number of extra-lessons. The unobserved term ϵ_Y stands for student's academic skills that cause both grades and college enrollment. The abilities of teachers are uncorrelated with the ability of students as the schools were randomly selected. This justifies the independence between ϵ_T and ϵ_Y.

Consider another example of a policy maker who intends to reduce burglary activity in a neighborhood. She observes that homeowners in gated communities experience less criminal activity but also invest more in security cameras. The policy maker considers subsidizing the creation of gated communities, but wonders whether she should subsidize home security systems instead. Therefore she seeks to evaluate the share of the causal effect of gated communities on burglary activity that can be attributed to gated communities' security apparatus. Treatment T is a binary indicator that takes value 1 if a home buyer decides to live in a gated community and zero otherwise. The mediator M consists of a binary indicator whether the house has security cameras. Outcome Y stands for the criminal activity. The unobserved confounder ϵ_T denotes the unobserved preferences of the home buyer regarding safety, which affects his decision to live in a gated community as well as his expenditure on home security. The unobserved confounder ϵ_Y refers to the unobserved characteristics of the burglars, i.e. their sensitivity towards security cameras. Burglars decide on trespassing Y based on the observed house security M and whether the house is in gated community T. The owner's safety preferences are not observed by the burglar and thereby can influence the burglar's decision only through the observed variables T, M. Fluctuations of real state prices are a natural candidate for instrumental variables.

An application of the partially confounded model developed here can be found in Dippel et al. (2018). They decompose the total effect of import competition T on populist voting Y in Germany into a direct effect and an indirect effect that is mediated by labor market adjustments M to international trade. All causal effects can be identified with the same instrumental variable Z proposed by Autor et al. (2013), which relies on other countries' exposure to trade with low-wage countries.

5 Examining the Binary Model under Partial Confounding

This section investigates the partially confounded model in light of the Local Instrumental Variable Model (LIV) of Heckman and Vytlacil (1999). It examines the case in which the treatment T and the mediator M are binary variables that take values in $\{0, 1\}$. Heckman and Vytlacil (1999) show that the identification of treatment effects stems from the first derivative of the outcome expectation with respect to the propensity score. We show that mediation effects are retrieved by the second derivative of the expected outcome with respect to the same propensity score.

Our application builds upon the mediation model of Table 9 where error terms ν_t, ν_Y represent random variables and the observed variables are described by $T = f_T(Z, \nu_T)$, $M = f_M(T, \nu_T, \nu_Y)$, $Y = f_Y(T, M, \nu_Y, \epsilon)$. Error term ϵ prevents outcome Y from being deterministic when conditioned on T, M, ν_Y. Error terms ν_Y, ϵ may have an arbitrary statistical dependence. The IV exogeneity condition can be stated as:

$$Z \perp\!\!\!\perp (\nu_T, \nu_Y, \epsilon), \tag{22}$$

and the partial confounding assumption is given by:

$$\nu_T \perp\!\!\!\perp (\nu_Y, \epsilon). \tag{23}$$

The statistical relationships (22)–(23) imply a useful model property:

$$(Z, \nu_T, T) \perp\!\!\!\perp (Y(m, t), \nu_Y) \tag{24}$$

The relationship is due to the fact that Z, ν_T are jointly independent of ν_Y, ϵ by (22)–(23). In addition, 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).

The Treatment Equation

Our identification strategy builds upon the , who assume that the treatment equation can be expressed as an inequality that is separable in Z and ν_T :

$$T = f_T(Z, \nu_T) \equiv \mathbf{1}[\zeta(Z) \geq \phi(\nu_T)] \tag{25}$$

Separability assumption (25) implies that a change in the IV's value $z \to z'$ induces all participants toward or against the treatment choice. Vytlacil (2002) shows that the separability assumption is equivalent to the monotonicity assumption of Imbens and Angrist (1994), that is, $T_i(z) \geq T_i(z') \forall i \in \mathcal{I}$ or $T_i(z) \leq T_i(z') \forall\, i \in \mathcal{I}$. We assume that $\phi(\nu_T)$ in (25) is absolutely continuous. Thus its CDF $F_{\phi(\nu_T)}(\cdot)$ is strictly increasing and can be applied to both sides of the inequality in (25) without loss of information. This allows expressing the treatment equation as:

$$T = \mathbf{1}[P_T(Z) \geq U_T], \tag{26}$$

where $P_T(Z) = P(T = t_1|Z)$ is the propensity score for treatment choice T, and variable $U_T = F_{\phi(\nu_T)}(\phi(\nu_T))$ is uniformly distributed in $[0, 1]$. $P_T(Z)$ denotes the random variable generated by a transformation of the instrument Z. The statistical independence $Z \perp\!\!\!\perp \nu_T$ implies that $P_T(Z) \perp\!\!\!\perp U_T$. Under this notation, the exogeneity condition (22) can be restated as:

$$P_T(Z) \perp\!\!\!\perp (U_T, \nu_T, \nu_Y, \epsilon). \tag{27}$$

Evaluating the counterfactual Mediator

We seek to identify the distribution of the counterfactual mediator $M(t) = f_M(t, \nu_T, \nu_Y); t \in \{0, 1\}$. The identification follows the analysis in Vytlacil (2002). The exogeneity condition (27) implies that $(M(t), U_T) \perp\!\!\!\perp P_T(Z)$, which is useful to restate the expectation $E(M \cdot T|P_T(Z) = p)$ as:

$$\begin{aligned} E(M \cdot T|P_T(Z) = p) &= E(M \cdot \mathbf{1}[T = 1]|P_T(Z) = p) && (28) \\ &= E(M(1) \cdot \mathbf{1}[P_T(Z) \geq U_T]|P_T(Z) = p) \\ &= E(M(1) \cdot \mathbf{1}[p \geq U_T]) \text{ due to } P_T(Z) \perp\!\!\!\perp (U_T, M(t)) && (29) \end{aligned}$$

We can employ equation (29) to examine the difference of expectations in (30) for $p, p' \in (0, 1)$ such

that $p > p'$.

$$E(M \cdot T|P_T(Z) = p) - E(M \cdot T|P_T(Z) = p') = E(M(1) \cdot \mathbf{1}[p \geq U_T]) - E(M(1) \cdot \mathbf{1}[p' \geq U_T]) \quad (30)$$

$$= E(M(1) \cdot \mathbf{1}[p \geq U_T \geq p'])$$

$$= \frac{E(M(1)|p \geq U_T \geq p')}{P(p \geq U_T \geq p')}$$

$$= \frac{E(M(1)|p \geq U_T \geq p')}{p - p'} \text{ as } U_T \sim \text{unif}[0,1] \quad (31)$$

The identification of $E(M(1)|U = p)$ is obtained by applying the limit for $p' \to p$ in equation (31):

$$\frac{\partial E(M \cdot T|P_T(Z) = p)}{\partial p} = E(M(1)|U = p) = P(M(1) = 1|U = p). \quad (32)$$

Replacing T by $(1 - T)$ in equations (28)–(31) enables us to identify $E(M(0) = m_1|U = p)$ by:

$$\frac{\partial E(M \cdot (1 - T)|P_T(Z) = p)}{\partial p} = -E(M(0) = 1|U = p). \quad (33)$$

The identification of the distribution of $M(t); t \in \{0, 1\}$ conditional on $U = u \in [0, 1]$ requires $E(M \cdot \mathbf{1}[T = t]|P_T(Z) = p)$ to be differentiable with respect to $p \in [0, 1]$. This means that the instrument Z must ensure enough variation around $P_T(Z) = p$. If Z ensures enough variation over its full support $[0, 1]$, then $E(M(t)|U_T = u)$ is identified for all $u \in [0, 1]$ and $E(M(t))$ is obtained by integrating $E(M(t)|U_T = u)$ over the unity interval.

The Mediation Equation

We assume that the mediator M is also described by a separable equation on T, U_T,

$$M = f_M(T, \nu_T, \nu_Y) \equiv \mathbf{1}[\xi(T, U_T) \geq \varphi(\nu_Y)], \quad (34)$$

where $U_T = F_{\phi(\nu_T)}(\phi(\nu_T))$. We assume that $\varphi(\nu_Y)$ is absolutely continuous and $\xi(t, u)$ is invertible in u, either strictly increasing or decreasing in u. Let $F_{\varphi(\nu_Y)}(\cdot)$ be the CDF of $\varphi(\nu_Y)$. We can apply this CDF transformation to both sides of the inequality in (34) to express the mediation equation

as:

$$M = \mathbf{1}[\tau(T, U_T) \geq U_Y], \tag{35}$$

where $U_Y = F_{\varphi(\nu_Y)}(\varphi(\nu_Y)) \sim \text{unif}[0,1]$ and $\tau(T, U_T) = F_{\varphi(\nu_Y)}(\xi(T, U_T))$. The independence $(T, Z) \perp\!\!\!\perp \nu_Y$ in (24) implies that $(T, Z) \perp\!\!\!\perp U_Y$. This relationship is useful to assess $\tau(T, U_T)$:

$$\begin{aligned}
P(M = 1|T = 1, P_T(Z) = p) &= E(M|T = 1, P_T(Z) = p) \\
&= E(\mathbf{1}[\tau(t_1, U_T) \geq U_Y]|T = t_1, P_T(Z) = p) \\
&= E(\mathbf{1}[\tau(t_1, p) \geq U_Y]|T = t_1, P_T(Z) = p) \\
&= E(\mathbf{1}[\tau(t_1, p) \geq U_Y]) \text{ due to } (T, Z) \perp\!\!\!\perp U_Y \\
&= E(\mathbf{1}[\tau(t_1, p) \geq U_Y]) = \tau(t_1, p) \text{ as } U_Y \sim \text{unif}[0,1]
\end{aligned} \tag{36}$$

Equation (36) means that the random variable $\tau(t, U_T)$ is a transformation of the instrument Z that stands for the propensity score of the mediator M conditioned on $T = t; t \in \{0, 1\}$, that is, $\tau(t, U_T) = P(M = 1|T = t, P_T(Z))$.

The invertibility of $\tau(t, u)$ with respect to u implies a one-to-one relation between the propensity score for the choice T, that is $P_T(Z)$, and the conditional propensity score for the mediator, that is, $P(M = 1|T = t, P_T(Z))$. Let the support of $P_T(Z)$ be given by the interval $[\underline{p}_T, \bar{p}_T] \subset [0, 1]$. Then, if $\tau(t, u)$ is strictly increasing, the support for $P(M = 1|T = 1, P_T(Z))$ is given by $[\underline{p}^1_M, \bar{p}^1_M]$ where $\underline{p}^1_M = P(M = 1|T = 1, P_T(Z) = \underline{p}_T)$ and $\bar{p}^1_M = P(M = 1|T = 1, P_T(Z) = \bar{p}_T$.

Evaluating the counterfactual Outcome

According to (24), the partial confounding assumption implies that $T, U_T \perp\!\!\!\perp Y(m, t)|U_Y$. But conditioned on U_Y, M is a function of T and U_T. Therefore, we have that $(T, M) \perp\!\!\!\perp Y(m, t)|U_Y$. This statistical independence means that the counterfactual outcome is as good as random when conditioned U_Y, which plays the role of a matching variable we seek to control for.

A key identifying property of the partial confounding assumption is the independence $U_T \perp\!\!\!\perp (Y(m, t), U_Y)$ in (24). This property is employed in the identification of the mean and distribution of the counterfactual outcome $Y(t, m)$. We first replace $E(M \cdot T|P_T(Z) = p)$ in equation (30) by

$E(Y \cdot M \cdot T | P_T(Z) = p)$ to obtain the following result:

$$E(Y \cdot M \cdot T | P_T(Z) = p) - E(Y \cdot M \cdot T | P_T(Z) = p') = \frac{E(Y(1,1) \cdot M(1) | \mathbf{1}[p \geq U_T \geq p'])}{p' - p} \tag{37}$$

$$\therefore \frac{\partial E(Y \cdot M \cdot T | P_T(Z) = p)}{\partial p} = E(Y(1,1) \cdot M(1) | U_T = p) \tag{38}$$

We can replace the $M(1)$ on the right-hand side of (38) by the mediation equation $M(t) = \mathbf{1}[\tau(t, U_T) \geq U_Y]; t \in \{0,1\}$ in (35):

$$\begin{aligned} E(Y(1,1) \cdot M(1) | U_T = p) &= E(Y(1,1) \cdot \mathbf{1}[\tau(1, U_T) \geq U_Y] | U_T = p) \\ &= E(Y(1,1) \cdot \mathbf{1}[\tau(1,p) \geq U_Y] | U_T = p) \\ &= E(Y(1,1) \cdot \mathbf{1}[\tau(1,p) \geq U_Y]), \end{aligned} \tag{39}$$

where the last equation (39) is due to the partial confounding property $U_T \perp\!\!\!\perp (Y(m,t), U_Y)$. Function $\tau(t,u)$ is strictly increasing in u, therefore we can apply the same differentiation rationale of (30)–(32) and (37)–(38) to equation (39):

$$E(Y(1,1) \cdot M(1) | U_T = p) - E(Y(1,1) \cdot M(1) | U_T = p') = \frac{E(Y(1,1) \cdot M(1) | \mathbf{1}[\tau(1,p) \geq U_Y \geq \tau(1,p')])}{p' - p}$$

$$\therefore \frac{\partial E(Y(1,1) \cdot M(1) | U_T = p)}{\partial p} = E(Y(1,1) \cdot M(1) | U_Y = \tau(1,p)). \tag{40}$$

We can now combine (38) and (40) to state the following result:

$$\frac{\partial^2 E(Y \cdot M \cdot T | P_T(Z) = p)}{\partial p^2} = \frac{\partial E(Y(1,1) \cdot M(1) \cdot T | U_T = p)}{\partial p} = E(Y(m_1, t_1) | U_Y = u), \tag{41}$$

$$\text{such that } u = P(M = 1 | T = 1, P_T(Z) = p). \tag{42}$$

Equation (43) states the general case that comprises the four cases where the treatment and the mediator takes value in $(m,t) \in \{0,1\} \times \{0,1\}$:

$$\frac{\partial^2 E(Y \cdot \mathbf{1}[M = m] \cdot \mathbf{1}[T = t] | P_T(Z) = p)}{\partial p^2} = (-1)^{1-m} \cdot (-1)^{1-t} \cdot E(Y(t,m) | U_Y = u); \tag{43}$$

$$\text{such that } u = P(M = 1 | T = t, P_T(Z) = p), t \in \{0,1\}.$$

In particular, we can define the Marginal Mediator Effect as $\Delta_M(t,u) = E(Y(t,1) - Y(t,0)|U_Y = u)$ which is identified as:

$$\Delta_M(t,u) = (-1)^{1-t} \frac{\partial^2 E(Y \cdot \mathbf{1}[T=t]|P_T(Z)=p)}{\partial p^2}, \tag{44}$$

for $t \in \{0,1\}$ and p such that $P(M=1|T=t, P_T(Z)=p) = u$.

The identification of causal parameters depends on the support of propensity scores $P_T(Z)$ and $P(M=1|T=t, P_T(Z)); t \in \{0,1\}$. Under full support, we can obtain the expected value the counterfactual mediator and the outcome as $E(M(t)) = \int_0^1 E(M(t)|U_T = u)du$ and $E(Y(m,t)) = \int_0^1 E(Y(m,t)|U_Y = u)du$ where $E(M(t)|U_T = u)$ is identified by (32) and $E(Y(m,t)|U_Y = u)$ by (43). The direct and indirect effects require the evaluation of $E(Y(t, M(t'))$ for $t, t' \in \{0,1\}$. This parameter is retrieved by:

$$E(Y(t,M(t')) = \int E(Y(t,m))dF_{M(t)}(m) = E(Y(t,0))(1 - E(M(t'))) + E(Y(t,1))E(M(t')). \tag{45}$$

6 Examining the Linear IV Model under Partial Confounding

This section investigates the linear model in which the equations of the partially confounded IV model of Table 9 are described by linear functions. We show that, under linearity, the mediation effects can be evaluated by standard Two Stage Least Square (2SLS) regressions.

Consider the model (46)–(50) and, for sake of notational simplicity, let the expected value of all the variables be zero.[18]

$$Z = \epsilon_Z, \tag{46}$$
$$T = \beta_T^Z \cdot Z + \epsilon_T, \tag{47}$$
$$M = \beta_M^T \cdot T + \epsilon_M, \tag{48}$$
$$Y = \beta_Y^T \cdot T + \beta_Y^M \cdot M + \epsilon_Y, \tag{49}$$
$$\epsilon_Z \perp\!\!\!\perp (\epsilon_T, \epsilon_M, \epsilon_Y) \tag{50}$$

[18]This is not a necessary nor binding requirement, the results of this section also hold if we relax this condition.

The statistical independence in (50) characterizes Z as an instrument. In the mediation nomenclature, β_Y^T is the direct effect of T on Y, the multiplication $\beta_Y^M \beta_M^T$ is the indirect effect of T on Y, and the sum of these effects, $\beta_{IV} = \beta_Y^T + \beta_Y^M \beta_M^T$ is the total effect of T on Y.

We are particularly interested in employing the instrument Z to evaluate the parameter β_Y^M which is the causal effect of M on Y. This parameter is necessary for decomposing the total effect on T on Y into its direct and its indirect effect. As discussed earlier, β_Y^M cannot be identified without further assumptions. Its identification stems from the partial confounding assumption (51) which states that error terms ϵ_T and ϵ_Y are statistically independent:

$$\text{Partial Confounding Assumption: } \epsilon_T \perp\!\!\!\perp \epsilon_Y \tag{51}$$

Assumption (51) does imply that error terms ϵ_T, ϵ_Y are statistically independent from ϵ_M. For instance, ϵ_M can be a mixture of error terms ϵ_T and ϵ_Y such as $\epsilon_M = \epsilon + \alpha \cdot \epsilon_T + (1-\alpha)\epsilon_Y$.

The identification of the model coefficients depends on the relation between the covariance of the observed variables and the covariance of the error terms. It is useful to represent equations (46)–(49) in matrix form. Let $\mathbf{X} = [Z, T, M, Y]'$ be the vector of observed random variables and $\epsilon = [\epsilon_Z, \epsilon_T, \epsilon_M, \epsilon_Y]'$ be the vector of unobserved error terms. Matrix $\mathbf{\Psi}$ in (52) stands for the arrangement of linear coefficients.

$$\underbrace{\begin{bmatrix} Z \\ T \\ M \\ Y \end{bmatrix}}_{\mathbf{X}} = \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ \beta_T^Z & 0 & 0 & 0 \\ 0 & \beta_M^T & 0 & 0 \\ 0 & \beta_Y^T & \beta_Y^M & 0 \end{bmatrix}}_{\mathbf{\Psi}} \cdot \underbrace{\begin{bmatrix} Z \\ T \\ M \\ Y \end{bmatrix}}_{\mathbf{X}} + \underbrace{\begin{bmatrix} \epsilon_Z \\ \epsilon_T \\ \epsilon_M \\ \epsilon_Y \end{bmatrix}}_{\epsilon}. \tag{52}$$

Equations (46)–(49) are thus written as $\mathbf{X} = \mathbf{\Psi} \cdot \mathbf{X} + \boldsymbol{\epsilon}$. Let $\Sigma_{\mathbf{X}}$ in (53) and Σ_ϵ in (54) be the covariance matrix of observed data and the error terms respectively. The zeros in Σ_ϵ are due to the statistical independence in (50) and (51).

$$\Sigma_{\mathbf{X}} \equiv \mathbf{Var}\begin{pmatrix} Z \\ T \\ M \\ Y \end{pmatrix} = \begin{bmatrix} \sigma_{ZZ} & \sigma_{ZT} & \sigma_{ZM} & \sigma_{ZY} \\ \cdot & \sigma_{TT} & \sigma_{TM} & \sigma_{TY} \\ \cdot & \cdot & \sigma_{MM} & \sigma_{MY} \\ \cdot & \cdot & \cdot & \sigma_{YY} \end{bmatrix}. \tag{53}$$

$$\Sigma_\epsilon \equiv \mathbf{Var}\begin{pmatrix} \epsilon_Z \\ \epsilon_T \\ \epsilon_M \\ \epsilon_Y \end{pmatrix} = \begin{bmatrix} \sigma^2_{\epsilon_Z} & 0 & 0 & 0 \\ \cdot & \sigma^2_{\epsilon_T} & \sigma_{\epsilon_T,\epsilon_M} & 0 \\ \cdot & \cdot & \sigma^2_{\epsilon_M} & \sigma_{\epsilon_M,\epsilon_Y} \\ \cdot & \cdot & \cdot & \sigma^2_{\epsilon_Y} \end{bmatrix}. \tag{54}$$

The identification of the model coefficients is based on the equality $(\mathbf{I}-\mathbf{\Psi})\,\Sigma_{\mathbf{X}}\,(\mathbf{I}-\mathbf{\Psi})' = \Sigma_{\boldsymbol{\epsilon}}$, which generates the following identification formulas:

$$\beta^T_M = \frac{\sigma_{ZM}}{\sigma_{ZT}} \tag{55}$$

$$\beta^M_Y = \frac{\sigma_{ZT}\sigma_{TY} - \sigma_{TT}\sigma_{ZY}}{\sigma_{ZT}\sigma_{TM} - \sigma_{TT}\sigma_{ZM}} \tag{56}$$

$$\beta^T_Y = -\frac{\sigma_{ZM}\sigma_{TY} - \sigma_{TM}\sigma_{ZY}}{\sigma_{ZT}\sigma_{TM} - \sigma_{TT}\sigma_{ZM}} \tag{57}$$

Each identifying formula is associated with a 2SLS estimator. Parameter β^T_M in (55) can be estimated by a 2SLS regression where Z is the IV, T is the endogenous explanatory variable, and M is the outcome variable as in (58)–(59). In essence, this 2SLS regression exploits exogeneity condition $Z \perp\!\!\!\perp M(t)$.

$$\text{First Stage:} \quad T = \beta^Z_T \cdot Z + \epsilon_T, \tag{58}$$

$$\text{Second Stage:} \quad M = \beta^T_M \cdot \hat{T} + \epsilon_M. \tag{59}$$

The estimation of parameters β^M_Y, β^T_Y in (56)–(57) exploits the fact that $Z \perp\!\!\!\perp Y(m)|T$. The parameters can be jointly estimated by a 2SLS regression where T plays the role of a conditioning variable, Z is the instrument, M is the endogenous variable, and Y is the dependent variable:

$$\text{First Stage:} \quad M = \gamma^Z_M \cdot Z + \gamma^T_M \cdot T + \epsilon_M, \tag{60}$$

$$\text{Second Stage:} \quad Y = \beta^M_Y \cdot \hat{M} + \beta^T_Y \cdot T + \epsilon_Y. \tag{61}$$

The estimate for the direct effect is given by $\hat{\beta}^T_Y$ in the 2SLS regression (60)–(61), the indirect effect estimate is given by $\hat{\beta}^T_M\hat{\beta}^M_Y$ and employs results from both 2SLS regressions described in (58)–(59) and (60)–(61). The total effect of T on Y can be estimated by the sum $\hat{\beta}^T_Y + \hat{\beta}^T_M\hat{\beta}^M_Y$. The total effect could also be estimated by using the exogeneity condition $Z \perp\!\!\!\perp Y(t)$, i.e. a 2SLS

regression that sets T as the endogenous variable and Z as the instrument:

$$\text{First Stage:} \quad T = \beta_T^Z \cdot Z + \epsilon_T, \tag{62}$$

$$\text{Second Stage:} \quad Y = \beta_{IV}^Y \cdot \hat{T} + \epsilon_Y. \tag{63}$$

For a just-identified regression using a single instrumental variable, the estimate $\hat{\beta}_{IV}^Y$ of the 2SLS (62)–(63) is numerically the same as the estimate $\hat{\beta}_Y^T + \hat{\beta}_M^T \hat{\beta}_Y^M$ based on the two previous 2SLS regressions. Indeed, if we compute $\beta_Y^T + \beta_M^T \beta_Y^M$ using the identification formulas (55)–(57), we obtain the well-known identification formula $\beta_{IV} = \frac{\sigma_{ZY}}{\sigma_{ZT}}$.

The linear model (46)–(50) is somewhat deceptive as the treatment variable T is statistically independent of outcome equations's unobserved term ϵ_Y. This is not a necessary condition to identify the mediation effects. One can consider a richer linear model (64)–(71) that includes unobserved mediators J, K.

$$Z = \epsilon_Z \tag{64}$$

$$T = \beta_T^Z \cdot Z + \epsilon_T \tag{65}$$

$$J = \beta_J^T \cdot T + \epsilon_J \qquad \text{Unobserved Pre-mediator} \tag{66}$$

$$M = \beta_M^T \cdot T + \beta_M^J \cdot J + \epsilon_M \tag{67}$$

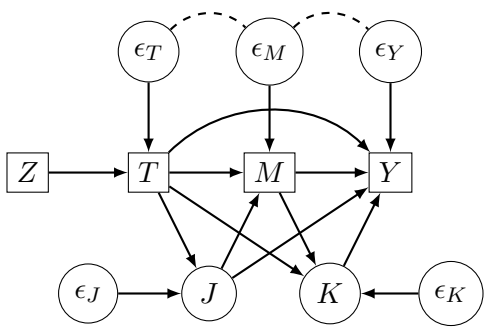
$$K = \beta_K^T \cdot T + \beta_K^M \cdot M + \epsilon_K \qquad \text{Unobserved Post-mediator} \tag{68}$$

$$Y = \beta_Y^T \cdot T + \beta_Y^M \cdot M + \beta_Y^J \cdot J + \beta_Y^K \cdot K + \epsilon_Y \tag{69}$$

$$\epsilon_Z \perp\!\!\!\perp (\epsilon_T, \epsilon_J, \epsilon_M, \epsilon_K, \epsilon_Y) \tag{70}$$

$$\epsilon_T \perp\!\!\!\perp (\epsilon_Y, \epsilon_K, \epsilon_J) \tag{71}$$

Model (64)–(71) is displayed as a DAG below:



The error terms $\epsilon_Y, \epsilon_K, \epsilon_J$ in Model (64)–(71) may have an arbitrary association. Yet, the

exogeneity conditions $Z \perp\!\!\!\perp Y(t), M(t)$ and $Z \perp\!\!\!\perp Y(m)|T$ hold. The independence relation in (71) characterizes the assumption of partial confoundedness. The treatment T is not independent of the unobserved terms in the outcome equation as it contains the unobserved mediators J, K that are caused by T. The direct effect in this model comprises the effect of T on Y via β_Y^T in addition to the impact that T has on Y via the unobserved mediators J, K. The causal effect of M on Y comprises the term β_Y^T in addition to the effect that M has on Y via K. The effect of T on M comprises the term β_M^T in (67) in addition to the effect that T has on M via J. Notationally, we have that:

$$\text{Effect of } T \text{ on } M: \quad \beta_M^T + \beta_M^J \cdot \beta_J^T \tag{72}$$

$$\text{Effect of } M \text{ on } Y: \quad \beta_Y^M + \beta_Y^K \cdot \beta_K^M \tag{73}$$

$$\text{Direct Effect of } T \text{ on } Y: \quad \beta_Y^T + \beta_Y^J \cdot \beta_J^T + \beta_Y^K \cdot \beta_K^T \tag{74}$$

$$\text{Indirect Effect of } T \text{ on } Y: \quad (\beta_Y^M + \beta_Y^K \cdot \beta_K^M) \cdot (\beta_Y^T + \beta_Y^J \cdot \beta_J^T + \beta_Y^K \cdot \beta_K^T) \tag{75}$$

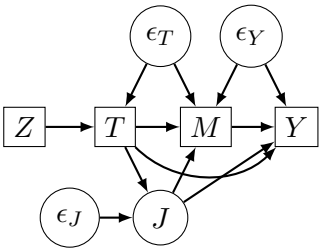
The effects described above are identified and can be estimated by the same 2SLS regressions that evaluate the parameters of the simpler linear model in (46)–(50). Specifically, the effect of M on Y in (73) is estimated by the 2SLS regression (58)–(59), while the effect of M on Y in (73) and the direct effect of T on Y in (74) are estimated by the 2SLS regression in (60)–(61).

7 Extensions of the Baseline Model

Last section suggests that the partially confounded model of Table 9 could be expressed in greater generality. Indeed, the direct causal link between T and Y may comprise an unobserved pre-mediator J that is caused by T and causes M, Y.[19] Table 10 lists the model equations.

[19]The causal direction plays a role in generating statistical associations. For instance, Z is unconditionally independent of ϵ_Y but it is not independent of J, that is, $Z \perp\!\!\!\perp \epsilon_Y$ and $Z \not\perp\!\!\!\perp J$. The opposite occurs when conditioning on T, that is, $Z \not\perp\!\!\!\perp \epsilon_Y|T$ and $Z \perp\!\!\!\perp J|T$.

Table 10: The Partially Confounded Model with Unobserved Pre-mediator J that causes Y

Model Equations	*DAG*	*Statistical Properties*
$Z = f_Z(\epsilon_Z)$ $T = f_T(Z, \epsilon_T)$ $J = f_J(T, \epsilon_J)$ $M = f_M(T, J, \epsilon_T, \epsilon_Y)$ $Y = f_Y(T, M, J, \epsilon_Y)$ $\epsilon_Z \perp\!\!\!\perp (\epsilon_T, \epsilon_J, \epsilon_Y)$ $\epsilon_T \perp\!\!\!\perp (\epsilon_Y, \epsilon_J)$		$Z \not\perp\!\!\!\perp T$ $Z \perp\!\!\!\perp (Y(t), M(t))$ $Z \not\perp\!\!\!\perp M \vert T$ $Z \perp\!\!\!\perp Y(m) \vert T$

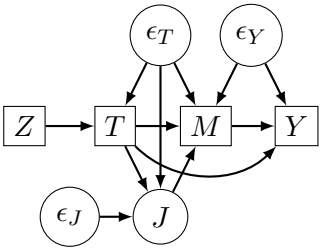
We can eliminate J by reiterated substitution and rewrite the equations of mediator M and outcome Y as: $M = g_M(T, \epsilon_T, \tilde{\epsilon}_Y)$ and $Y = g_Y(T, M, \tilde{\epsilon}_Y)$ where $\tilde{\epsilon}_Y = (\epsilon_Y, \epsilon_J)$. This means that the model in Table 10 is subsumed by the partial confounding model of Table 9.[20] As a consequence, the exogeneity condition $Z \perp\!\!\!\perp Y(m)|T$ holds.

The exogeneity condition $Z \perp\!\!\!\perp Y(m)|T$ would not hold if the error term ϵ_T caused J. If we replace equation $J = f_J(T, \epsilon_J)$ of Table 10 by $J = f_J(T, \epsilon_J, \epsilon_T)$, then ϵ_T would cause Y by a channel other than through T and M, namely $\epsilon_T \to J \to Y$. The iterated substitution would render an outcome equation that is a function of $T, M, \epsilon_T, \epsilon_Y$. The counterfactual outcome $Y(m)$ would take the following expression $Y(m) = g_Y(T, m, \epsilon_Y, \epsilon_T)$. In this case, conditioning on T would induce a correlation between Z and ϵ_T, which violates the statistical independence of Z and $Y(m)$.

We can regain the validity of the exogeneity condition $Z \perp\!\!\!\perp Y(m)|T$ if the pre-mediator U does not directly cause Y. Table 11 lists the equations of this model. In this case, we can eliminate J from the mediator equation $M = f_M(T, J, \epsilon_T, \epsilon_Y)$. The resulting equation can be expressed as $M = g_M(T, \tilde{\epsilon}_T, \epsilon_Y)$ where $\tilde{\epsilon}_T = (\epsilon_T, \epsilon_J)$ and the model becomes equivalent to the partially confounded model of Table 9.

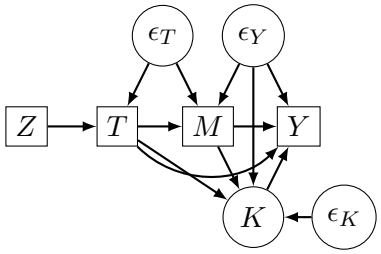
[20] The causal chain $T \to J \to M \to Y$ is contained in the indirect effect $T \to M \to Y$, while $T \to J \to Y$ is contained in the direct effect $T \to Y$.

Table 11: The Partially Confounded Model with Unobserved Pre-mediator U that does not cause Y

Model Equations	*DAG*	*Statistical Properties*
$Z = f_Z(\epsilon_Z)$ $T = f_T(Z, \epsilon_T)$ $U = f_J(T, \epsilon_T, \epsilon_J)$ $M = f_M(T, J, \epsilon_T, \epsilon_Y)$ $Y = f_Y(T, M, \epsilon_Y)$ $\epsilon_Z \perp\!\!\!\perp (\epsilon_T, \epsilon_J, \epsilon_Y)$ $(\epsilon_T, \epsilon_J) \perp\!\!\!\perp \epsilon_Y$		$Z \not\perp\!\!\!\perp T$ $Z \perp\!\!\!\perp (Y(t), M(t))$ $Z \not\perp\!\!\!\perp M \vert T$ $Z \perp\!\!\!\perp Y(m) \vert T$

The exogeneity condition $Z \perp\!\!\!\perp Y(m)|T$ still holds if we append the partially confounded model in Table 9 by a post-mediation variable that causes Y and is caused by confounding variable ϵ_Y. Table 12 lists the model equations. We can eliminate K from the outcome equation $Y = f_Y(T, M, K, \epsilon_Y)$ to obtain $Y = f_Y(T, M, \tilde{\epsilon}_Y)$ where $\tilde{\epsilon}_Y = (\epsilon_K, \epsilon_Y)$, which complies with the specifications of the original partially confounded model.

Table 12: The Partially Confounded Model with Unobserved Post-mediator K

Model Equations	*DAG*	*Statistical Properties*
$Z = f_Z(\epsilon_Z)$ $T = f_T(Z, \epsilon_T)$ $M = f_M(T, J, \epsilon_T, \epsilon_Y)$ $K = f_K(T, M, \epsilon_Y, \epsilon_K)$ $Y = f_Y(T, M, K, \epsilon_Y)$ $\epsilon_Z \perp\!\!\!\perp (\epsilon_T, \epsilon_K, \epsilon_Y)$ $\epsilon_T \perp\!\!\!\perp (\epsilon_K, \epsilon_Y)$		$Z \not\perp\!\!\!\perp T$ $Z \perp\!\!\!\perp (Y(t), M(t))$ $Z \not\perp\!\!\!\perp M \vert T$ $Z \perp\!\!\!\perp Y(m) \vert T$

8 Conclusions

This paper is motivated by a common inquiry among empirical economists that employ instrumental variables to estimate treatment effects: is it possible to exploit instrumental variables to beyond treatment effects and evaluate the causal mechanism among outcomes?

The question is empirically relevant as the typical IV model consist of a treatment variable T, a single or a limited number of instrumental variables Z, and numerous outcomes Y. It is often the case that the direction of causal relations among outcomes is known, for instance, an

intermediate outcome M causes a final outcome Y. A single instrumental variable can be used to identify the causal effect of one specific treatment on multiple outcomes. However, the standard IV model is not suited to identify causal effects among the outcomes. The exogeneity condition $Z \perp\!\!\!\perp (T(z), M(t), Y(t))$ that characterizes instrumental variable Z does not identify the causal effect of an intermediate outcome M on a final outcome Y.

This paper investigates conditions that enable the use of a given instrumental variable to non-parametrically identify the causal effect of intermediate outcome M on final outcome Y, while retaining the endogeneity properties of the treatment T with respect to the intermediate outcome M and final outcome Y.

The task of evaluating the causal chain $T \to M \to Y$ is often called mediation analysis. In the related literature, the intermediate outcome M is termed the mediator. Endogeneity stems from unobserved confounding variables that jointly cause T, M and Y. We examine possibilities to relax the statistical dependency among the unobserved confounding variables that cause T, M, Y in order to generate a new exogeneity condition that is useful to identify mediation effects without revoking the endogeneity of the treatment variable T with respect to the mediator M and outcome Y. We show that the only condition that satisfies all these criteria is the partial confounding assumption. That is to say that the confounding variables that jointly cause the treatment T and the mediator M are statistically independent of the confounding variables that jointly cause the mediator M and the final outcome Y.

The partial confounding assumption does not relax the endogeneity of the treatment T with respect to the mediator M or the final outcome Y. Thus, it does not modify nor impose additional assumptions when compared to the original IV model that motivates our analysis. Nevertheless, the partial confounding assumption generates a new exogeneity condition $Z \perp\!\!\!\perp Y(m)|T$ which states that the instrumental variable Z is statistically independent of the $Y(m)$ for a fixed mediator value m when conditioned on the treatment T. We show that the partially confounded model can be identified relying on well-known econometric methods, such as the control function approach, the linearity assumption, Two-stage Least squares, monotonicty and separability conditions.

References

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

Altonji, J. G. and R. L. Matzkin (2005, July). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–1102.

Attanasio, O., S. Cattan, E. Emla Fitzsimons, C. Meghir, and M. Rubio-Codina (2020). Estimating the production function for human capital: Results from a randomized controlled trial in colombia. *American Economic Review 110*, 48–85.

Autor, D. H., D. Dorn, and G. H. Hanson (2013). The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review 103*, 2121–2168.

Blundell, R. and J. Powell (2003). Endogeneity in nonparametric and semiparametric regression models. In L. P. H. M. Dewatripont and S. J. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Volume 2. Cambridge, UK: Cambridge University Press.

Blundell, R. and J. Powell (2004, July). Endogeneity in semiparametric binary response models. *Review of Economic Studies 71*(3), 655–679.

Brunello, G., M., N. S. Fort, and R. Winter-Ebmer (2016). The causal effect of education on health: What is the role of health behaviors? *Health Economics 25*, 314–336.

Chen, S. H., Y.-C. Chen, and J.-T. Liu (2019). The impact of family composition on educational achievement. *Journal of Human Resources 54*(1), 122–170.

Dippel, C., R. Gold, S. Heblich, and R. Pinto (2018). Instrumental variables and causal mechanisms: Unpacking the effect of trade on workers and voters. *NBER Working Paper* (w23209).

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

Flores, C. A. and A. Flores-Lagunes (2010). Nonparametric partial identi?cation of causal net and mechanism average treatment effects. *Umpublished Mimeo*.

Frölich, M., M. and M. Huber (2017). Direct and indirect treatment effects - causal chains and mediation analysis with instrumental variables. *Journal of the Royal Statistical Society B*. Online Version of Record published before inclusion in an issue.

Heckman, J. and R. Pinto (2017). Unordered monotonicity. *Forthcoming Econometrica*.

Heckman, J. J., R. Pinto, and P. A. Savelyev (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 2052–2086.

Heckman, J. J. and E. J. Vytlacil (1999, April). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*(8), 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2001). Local instrumental variables. In C. Hsiao, K. Morimune, and J. L. Powell (Eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, pp. 1–46. New York: Cambridge University Press.

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Imai, K., D. Tingley, and T. Yamamoto (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society A 176*(1), 5–51.

Imbens, G. W. and J. D. Angrist (1994, March). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and W. K. Newey (2007). Identification and estimation of triangular simultaneous equations models without additivity. Unpublished manuscript, Harvard University and MIT.

Imbens, G. W. and W. K. Newey (2009, September). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512.

Joffe, M. M., T. T. Small, D.and Have, S. Brunelli, and H. I. Feldman (2008). Extended instrumental variables estimation for overall effects. *International Journal of Biostatistics 4*(1).

Lee, S. and B. Salanié (2015). Identifying effects of multivalued treatments.

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

Matzkin, R. L. (2003, September). Nonparametric estimation of nonadditive random functions. *Econometrica 71*(5), 1339–1375.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*, 411–420.

Pearl, J. (2012). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. *Prevention Science 13*, 426–436.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods 19*, 459–481.

Pinto, R. (2015). Selection bias in a controlled experiment: The case of Moving to Opportunity. Unpublished Ph.D. Thesis, University of Chicago, Department of Economics.

Robins, J. M. and S. Greenland (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology 3*(2), 143–155.

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

Vytlacil, E. J. (2002, January). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Yamamoto, T. (2013). Identification and estimation of causal mediation effects with treatment noncompliance. *Unpublished Manuscript, MIT Department of Political Science*.

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