Econometric Reviews, 34(1–2):6–31, 2015

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474938.2014.944466

Taylor & Francis
Taylor & Francis Group

Econometric Mediation Analyses: Identifying the Sources of Treatment Effects from Experimentally Estimated Production Technologies with Unmeasured and Mismeasured Inputs

James J. Heckman and Rodrigo Pinto
Department of Economics, University of Chicago, Chicago, Illinois, USA

This paper presents an econometric mediation analysis. It considers identification of production functions and the sources of output effects (treatment effects) from experimental interventions when some inputs are mismeasured and others are entirely omitted.

Keywords Measurement error; Mediation analysis; Missing inputs; Production function.

JEL Classification D24; C21; C43; C38.

1. INTRODUCTION

William Barnett is a pioneer in the development and application of index theory and productivity accounting analyses. He has also pioneered the estimation of production functions. This paper follows in the tradition of Barnett's work. It develops an econometric mediation analysis to explain the sources of experimental treatment effects. It considers how to use experiments to identify production functions in the presence of unmeasured and mismeasured inputs. The goal of the analysis is to determine the causes of effects: sources of the treatment effects properly attributable to experimental variation in measured inputs.

Social experiments usually proceed by giving a vector of inputs to the treatment group and withholding it from the control group. Analysts of social experiments report a variety of treatment effects. Our goal is to go beyond the estimation of treatment effects and examine the mechanisms through which experiments generate these effects. Thus we seek to use experiments to estimate the production functions producing treatment effects. This exercise is called mediation analysis in the statistics literature (Imai et al., 2010, 2011; Pearl, 2011). Such analyses have been used for decades in economics (Klein and

Address correspondence to Rodrigo Pinto, Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, USA; E-mail: rodrig@uchicago.edu

Goldberger, 1955; Theil, 1958) and trace back to the work on path analysis by Sewall Wright (1934, 1921).

We provide an economically motivated interpretation of treatment effects. Treatment may affect outcomes through changing inputs. Treatment may also affect outcomes through shifting the map between inputs and outputs for treatment group members. When there are unmeasured (by the analyst) inputs, empirically distinguishing these two cases becomes problematic. We present a framework for making this distinction in the presence of unmeasured inputs and when the measured inputs are measured with error.

A fundamental problem of mediation analysis is that even though we might observe experimental variation in some inputs and outputs, the relationship between inputs and outputs might be confounded by unobserved variables. There may exist relevant unmeasured inputs changed by the experiment that impact outputs. If unmeasured inputs are not statistically independent of measured ones, then the observed empirical relation between measured inputs and outputs might be due to the confounding effect of experimentally induced changes in unmeasured inputs. In this case, treatment effects on outputs can be wrongly attributed to the enhancement of measured inputs instead of experimentally induced increase in unmeasured inputs.

Randomized Controlled Trials (RCTs) generate independent variation of treatment which allows the analyst to identify the causal effect of treatment on measured inputs and outputs. Nevertheless, RCTs unaided by additional assumptions do not allow the analyst to identify the causal effect of increases in measured inputs on outputs nor do they allow the analyst to distinguish between treatment effects arising from changes in production functions induced by the experiment or changes in unmeasured inputs when there is a common production function for treatments and controls.

This paper examines these confounding effects in mediation analysis. We demonstrate how econometric methods can be used to address them. We show how experimental variation can be used to increase the degree of confidence in the validity of the exogeneity assumptions needed to make valid causal statements. In particular, we show that we can test some of the strong assumptions implicitly invoked to infer causal effects in statistical mediation analyses. We analyze the invariance of our estimates of the sources of treatment effects to changes in measurement schemes.

The paper is organized in the following fashion. Section 2 discusses the previous literature and defines the mediation problem as currently framed in the statistics literature. Section 3 presents a mediation analysis within a linear framework with both omitted and mismeasured inputs. Subsection 4 discusses identification. Section 5 presents an estimation method. Subsection 6.1 discusses an invariance property when input measures are subject to affine transformations. Subsection 6.2 discusses further invariance results for general monotonic transformations of measures and for nonlinear technologies. Section 7 concludes.

2. ASSUMPTIONS IN STATISTICAL MEDIATION ANALYSIS

The goal of mediation analysis as framed in the literature in statistics is to disentangle the average treatment effect on outputs that operates through two channels: (1) *indirect* output effects arising from the effect of treatment on measured inputs and (2) *direct* output effects that operate through channels other than changes in the measured inputs. The mediation literature often ignores the point that Direct Effects are subject to some ambiguity: they can arise from inputs changed by the experiment that are not observed by the analyst, but can also arise from changes in the map between inputs and outputs.

To clarify ideas, it is useful to introduce some general notation. Let D denote treatment assignment. $D = 1$, if an agent is treated, and $D = 0$ otherwise. Let Y_1 and Y_0 be counterfactual outputs when D is *fixed* at "1" and "0," respectively. By fixing, we mean an independent manipulation where treatment status is set at d. The distinction between fixing and conditioning traces back to Haavelmo (1943). For recent discussions see Pearl (2011, 2001) and Heckman and Pinto (2012). We use the subscript $d \in \{0, 1\}$ to represent variables when treatment is fixed at d. In this notation, Y_d represents output Y when treatment status is fixed at d, and the realized output is given by

$$Y = DY_1 + (1 - D)Y_0. \tag{1}$$

In our notation, the average treatment effect between treatment and control groups is given by

$$ATE = E(Y_1 - Y_0). \tag{2}$$

We define a vector of inputs when treatment is *fixed* at d by $\boldsymbol{\theta}_d = (\theta_d^j : j \in \mathcal{J})$, where $\mathcal{J}$ is an index set for inputs. We define the vector of realized inputs by $\boldsymbol{\theta}$ in a fashion analogous to Y: $\boldsymbol{\theta} = D\boldsymbol{\theta}_1 + (1 - D)\boldsymbol{\theta}_0$. While output Y is assumed to be observed, we allow for some inputs to be unobserved. Notationally, let $\mathcal{J}_p \subseteq \mathcal{J}$ be the index set of *proxied inputs*—inputs for which we have observed measurements. We represent the vector of proxied inputs by $\boldsymbol{\theta}_d^p = (\theta_d^j : j \in \mathcal{J}_p)$. We allow for the possibility that observed measurements may be imperfect proxies of measured inputs so that measured inputs may not be observed directly. We denote the remaining inputs indexed by $\mathcal{J} \backslash \mathcal{J}_p$ as *unmeasured inputs*, which are represented by $\boldsymbol{\theta}_d^u = (\theta_d^j : j \in \mathcal{J} \backslash \mathcal{J}_p)$.

We postulate that the output Y is generated through a production function whose arguments are both measured and unmeasured inputs in addition to an auxiliary set of baseline variables $\boldsymbol{X}$. Variables in $\boldsymbol{X}$ are assumed not to be caused by treatment D that affects output Y in either treatment state. The production function for each treatment regime is

$$Y_d = f_d(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_d^u, \boldsymbol{X}), d \in \{0, 1\}. \tag{3}$$

Equation (3) states that output Y_d under treatment regime $D = d$ is generated by $(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_d^u, \boldsymbol{X})$ according to function f_d such that $d \in \{0, 1\}$. If $f_1 = f_0$, functions (f_1, f_0) are said to be invariant across treatment regimes. Invariance means that the relationship between inputs and output is determined by a stable mechanism defined by a deterministic function unaffected by treatment.

From Eq. (2), the average treatment effect or *ATE* is given by

$$ATE = E(f_1(\boldsymbol{\theta}_1^p, \boldsymbol{\theta}_1^u, \boldsymbol{X}) - f_0(\boldsymbol{\theta}_0^p, \boldsymbol{\theta}_0^u, \boldsymbol{X})).$$

Expectations are computed with respect to all inputs. Treatment effects operate through the impact of treatment D on inputs $(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_d^u), d\epsilon\{0, 1\}$ and also by changing the map between inputs and the outcome, namely, $f_d(\cdot); d \in \{0, 1\}$. Observed output is given by $Y = \sum_{d\in\{0,1\}} \mathbf{1}[D = d] \cdot f_d(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_d^u, \boldsymbol{X})$.

We are now equipped to define mediation effects. Let $Y_{d,\bar{\boldsymbol{\theta}}_d^p}$ represent the counterfactual output when treatment status D is fixed at d and proxied inputs are fixed at the some value $\bar{\boldsymbol{\theta}}_d^p \in \operatorname{supp} \boldsymbol{\theta}_{\boldsymbol{d}}^{\boldsymbol{p}}$. From production function (3),

$$Y_{d,\bar{\boldsymbol{\theta}}^p} = f_d(\bar{\boldsymbol{\theta}}_d^p, \boldsymbol{\theta}_d^u, \boldsymbol{X}), d \in \{0, 1\}. \tag{4}$$

Note that the subscript d of $Y_{d,\bar{\boldsymbol{\theta}}^p}$ arises both from the selection of the production function $f_d(\cdot)$, from the choice of d, and from changes in unmeasured inputs $\boldsymbol{\theta}_d^u$. Moreover, conditional on $\boldsymbol{X}$ and fixing $\boldsymbol{\theta}_d^p = \bar{\boldsymbol{\theta}}_d^p$, the source of variation of $Y_{d,\bar{\boldsymbol{\theta}}^p}$ is attributable to unmeasured inputs $\boldsymbol{\theta}_d^u$. Keeping $\boldsymbol{X}$ implicit, use $Y_{d,\bar{\boldsymbol{\theta}}_{d'}^p}$ to represent the value output would take fixing D to d and simultaneously fixing measured inputs to be $\bar{\boldsymbol{\theta}}_{d'}^p$.

In the mediation literature, total treatment effect is called the ATE. It is often decomposed into *direct* and *indirect* treatment effects. The indirect effect (IE) is the effect of changes in the distribution of proxied inputs (from $\boldsymbol{\theta}_0^p$ to $\boldsymbol{\theta}_1^p$) on mean outcomes while holding the technology f_d and the *distribution* of unmeasured inputs $\boldsymbol{\theta}_d^u$ fixed at treatment status d. Formally, the indirect effect is

$$\begin{aligned} IE(d) &= E(Y_d(\boldsymbol{\theta}_1^p) - Y_d(\boldsymbol{\theta}_0^p)) \\ &= E(f_d(\boldsymbol{\theta}_1^p, \boldsymbol{\theta}_d^u, \boldsymbol{X}) - f_d(\boldsymbol{\theta}_0^p, \boldsymbol{\theta}_d^u, \boldsymbol{X})). \end{aligned}$$

Here expectations are taken with respect to $\boldsymbol{\theta}_d^u$ and $\boldsymbol{X}$. One definition of the direct effect (DE) is the average effect of treatment holding *measured* inputs fixed at the level appropriate to treatment status d but allowing technologies and associated distributions of unobservables to change with treatment regime:

$$\begin{aligned} DE(d) &= E(Y_{1,\boldsymbol{\theta}_d^p} - Y_{0,\boldsymbol{\theta}_d^p}) \\ &= E(f_1(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_1^u, \boldsymbol{X}) - f_0(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_0^u, \boldsymbol{X})). \end{aligned} \tag{5}$$

Robin (2003) terms these effects as the pure, direct, and indirect effects, while Pearl (2001) calls them the natural direct and indirect effects.

We can further decompose the direct effect of Eq. (5) into portions associated with the change in the distribution of $\boldsymbol{\theta}_d^u; d \in \{0,1\}$ and the change in the map between inputs and outputs $f_d(\cdot); d \in \{0,1\}$. Define

$$DE'(d,d') = E(f_1(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_{d'}^u, \boldsymbol{X}) - f_0(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_{d'}^u, \boldsymbol{X})). \tag{6}$$

$DE'(d,d')$ is the treatment effect mediated by changes in the map between inputs and outputs when fixing the distribution of measured inputs at $\boldsymbol{\theta}_d^p$, and unmeasured inputs at $\boldsymbol{\theta}_{d'}^u$ for $d,d' \in \{0,1\}$. Define

$$DE''(d,d') = E(f_{d'}(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_1^u, \boldsymbol{X}) - f_{d'}(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_0^u, \boldsymbol{X})). \tag{7}$$

$DE''(d,d')$ is the treatment effect mediated by changes unmeasured inputs from $\boldsymbol{\theta}_0^u$ to $\boldsymbol{\theta}_1^u$ while setting the production function at $f_{d'}(\cdot)$ where measured inputs are fixed at $\boldsymbol{\theta}_d^p$ for $d,d' \in \{0,1\}$.

The definition of direct effect in Eq. (5) is implicit in the mediation literature. Definitions (6) and (7) are logically coherent. Direct effects (5) can be written alternatively as:

$$DE(d) = DE'(d,1) + DE''(d,0);$$
$$DE(d) = DE'(d,0) + DE''(d,1).$$

The source of the direct treatment effect is often ignored in the statistical literature. It can arise from changes in unobserved inputs induced by the experiment (from $\boldsymbol{\theta}_0^u$ to $\boldsymbol{\theta}_1^u$). It could also arise from an empowerment effect, e.g., that treatment modifies the technology that maps inputs into outputs (from f_0 to f_1). The change in technology may arise from new inputs never previously available such as parenting information as studied by Cunha (2012) and Heckman et al. (2012). If both measured and unmeasured inputs were known (including any new inputs never previously available to the agent), then the causal relationship between inputs and outputs could be estimated. Using the production function for each treatment state, one could decompose treatment effects into components associated with changes in either measured or unmeasured inputs. Since unmeasured inputs are not observed, the estimated relationship between measured inputs and outputs may be confounded with changes in unmeasured inputs induced by the experiment.

In this framework, under the definition of a direct effect (Eq. 5), we can decompose the total treatment effect into the direct and indirect effect as follows:

$$ATE = E(Y_1(\boldsymbol{\theta}_1^p) - Y_0(\boldsymbol{\theta}_0^p))$$

$$= \underbrace{E(Y_1(\boldsymbol{\theta}_1^p) - Y_0(\boldsymbol{\theta}_1^p))}_{\text{DE}(1)} + \underbrace{E(Y_0(\boldsymbol{\theta}_1^p) - Y_0(\boldsymbol{\theta}_0^p))}_{\text{IE}(0)}$$
$$= \underbrace{E(Y_1(\boldsymbol{\theta}_0^p) - Y_0(\boldsymbol{\theta}_0^p))}_{\text{DE}(0)} + \underbrace{E(Y_1(\boldsymbol{\theta}_1^p) - Y_1(\boldsymbol{\theta}_0^p))}_{\text{IE}(1)}.$$

The literature of mediation analysis deals with the problem of confounding effects of unobserved inputs and the potential technology changes by invoking different assumptions. We now examine those assumptions.

The standard literature on mediation analysis in psychology regresses outputs on mediator inputs (Baron and Kenny, 1986). The assumptions required to give these regressions a causal interpretation are usually not explicitly stated. This approach often adopts the strong assumption of *no variation* in unmeasured inputs conditional on the treatment. Under this assumption, measured and unmeasured inputs are statistically independent. Moreover, the effect of unmeasured inputs $\boldsymbol{\theta}_d^u$ are fully summarized by a dummy variable for treatment status. In addition, this literature assumes full invariance of the production function, that is, $f_1(\cdot) = f_0(\cdot)$. Under these assumptions, function (3) reduces to

$$Y_d = f(\boldsymbol{\theta}_d^p, d, \boldsymbol{X}), \tag{8}$$

which can readily be identified and estimated. A similar framework is also used by in Pearl (2001).

Imai et al. (2011) and Imai et al. (2010) present a different analysis and invoke two conditions in their *Sequential Ignorability Assumption*. Their approach does not explicitly account for unobserved inputs. They invoke statistical relationships that can be interpreted as a double randomization, i.e., they assume that both treatment status and measured inputs are randomized. More specifically, their approach assumes independence of both treatment status D and measured inputs $\boldsymbol{\theta}_d^p$ with respect to $Y_{d,\bar{\boldsymbol{\theta}}_d^p}$ conditional on covariates $\boldsymbol{X}$.

Assumption A-1. Sequential Ignorability (Imai et al., 2010, 2011).

(i) $(Y_{d,\bar{\boldsymbol{\theta}}_d^p}, \boldsymbol{\theta}_{d'}^p) \perp\!\!\!\perp D \mid \boldsymbol{X};\ d, d' \in \{0,1\}$
(ii) $Y_{d,\bar{\boldsymbol{\theta}}_d^p} \perp\!\!\!\perp \boldsymbol{\theta}_{d'}^p \mid D, \boldsymbol{X};\ d, d' \in \{0,1\}$
(iii) $0 < Pr(D = d \mid \boldsymbol{X}) < 1$ and $0 < Pr(\boldsymbol{\theta}_d^p = \boldsymbol{\theta} \mid D = d, \boldsymbol{X}) < 1;\ d \in \{0,1\}, \forall \boldsymbol{\theta}_d^p, \boldsymbol{\theta} \in \text{supp}(\boldsymbol{\theta}_d^p)$.

Condition (i) of Assumption (A-1) states that both counterfactual outputs and counterfactual measured inputs are independent of D conditional on preprogram variables. These statistical relationships are generated by a RCT that randomly assigns treatment status D given $\boldsymbol{X}$. Indeed, if treatment status D were randomly assigned by a randomization protocol that conditions on preprogram variables $\boldsymbol{X}$, then $Y_d \perp\!\!\!\perp D \mid \boldsymbol{X}$ (see,

e.g., Heckman et al., 2010, for a discussion). But proxied and unmeasured inputs are also outcomes in a RCT and, therefore, $(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_d^u) \perp\!\!\!\perp D \mid \boldsymbol{X}$. Condition (i) of Assumption (A-1) is invoked to eliminate the dependence arising from the fact that for fixed $\boldsymbol{X}$ the source of variation of $Y_{d,\bar{\boldsymbol{\theta}}_d^p}$ is $\boldsymbol{\theta}_d^u$.

Condition (ii) declares that counterfactual outcomes given d and $\bar{\boldsymbol{\theta}}_d^p$ are independent of unmeasured inputs given the observed treatment status and the preprogram variables $\boldsymbol{X}$. In other words, input $\boldsymbol{\theta}_{d'}^p$ is statistically independent of *potential* outputs when treatment is fixed at $D = d$ and measured inputs are fixed at $\bar{\boldsymbol{\theta}}_{d'}^p$ conditional on treatment assignment D and same preprogram characteristics $\boldsymbol{X}$. The same randomization rationale used to interpret Condition (i) can be applied to Condition (ii). Thus Condition (ii) can be understood as if a second RCT were implemented for each treatment group such that measured inputs are randomized through a randomization protocol conditional on preprogram variables $\boldsymbol{X}$ and treatment status D. This randomization is equivalent to assuming that $\boldsymbol{\theta}_d^p \perp\!\!\!\perp \boldsymbol{\theta}_{d'}^0$ for all d and d'. Condition (iii) is a support condition that allows the estimation of treatment effects conditioned on the values $\boldsymbol{X}$ takes. Even though the Imai et al. (2010) and Imai et al. (2011) approach is weaker than the Pearl (2001) solution which is based on lack of variation of unobserved inputs, their assumptions are nonetheless still quite strong.

Imai et al. (2010) show that under Assumption A-1, the direct and indirect effects are given by

$$E(IE(d) \mid \boldsymbol{X}) = \int E(Y \mid \boldsymbol{\theta}^p = t, D = d, \boldsymbol{X})(dF_{(\boldsymbol{\theta}^p \mid D=1,\boldsymbol{X})}(t) - dF_{(\boldsymbol{\theta}^p \mid D=0,\boldsymbol{X})}(t)) \tag{9}$$

$$E(DE(d) \mid \boldsymbol{X}) = \int (E(Y \mid \boldsymbol{\theta}^p = t, D = 1, \boldsymbol{X}) - E(Y \mid \boldsymbol{\theta}^p = t, D = 0, \boldsymbol{X}))dF_{(\boldsymbol{\theta}^p \mid D=1,\boldsymbol{X})}(t). \tag{10}$$

Pearl (2011) uses the term *Mediation Formulas* for Eqs. (9) and (10). Like Imai et al. (2010), Pearl (2011) invokes the assumption of exogeneity on mediators conditioned on variables $\boldsymbol{X}$ to generate these equations.

Identification of the direct and indirect effects under the strong implicit assumption A-1, translates to an assumption of no-confounding effects on both treatment and measured inputs. This assumption does not follow from a randomized assignment of treatment. Randomized trials ensure independence between treatment status and counterfactual inputs/outputs, namely $Y_d \perp\!\!\!\perp D \mid \boldsymbol{X}$ and $\boldsymbol{\theta}_d^p \perp\!\!\!\perp D \mid \boldsymbol{X}$. Thus RCTs identify treatment effects for proxied inputs and for outputs. However, random treatment assignment does not imply independent variation between proxied inputs $\boldsymbol{\theta}_d^p$ and unmeasured inputs $\boldsymbol{\theta}_d^u$. In particular, it does not guarantee independence between counterfactual outputs $Y_{d,\bar{\boldsymbol{\theta}}_d^p}$, which is generated in part by $\boldsymbol{\theta}_d^u$, and measured inputs $\boldsymbol{\theta}_{d'}^p$ as assumed in Condition (ii) of Assumption A-1.

2.1. Mediation Analysis under RCT

It is useful to clarify the strong causal relationships implied by Condition (ii) of Assumption A-1 in light of a mediation model based on a RCT. To this end, we first start by defining a standard confounding model arising from uncontrolled preprogram unobserved variables. We then introduce a general RCT model and establish the benefits of RCTs in comparison with models that rely on standard matching assumptions. We then define a general mediation model with explicitly formulated measured and unmeasured inputs. We then examine the causal relationships of the mediation model that are implied by Condition (ii) of Assumption A-1. We show that the assumptions made in Assumption A-1 are stronger than standard assumptions invoked in matching.

A standard confounding model can be represented by the following three variables: (1) an output of interest Y; (2) a treatment indicator D that causes the output of interest. As before, we use $D = 1$ for treated and $D = 0$ for untreated; and (3) an unobserved variable $\boldsymbol{V}$ that causes both D and Y. A major difference between unobserved variable $\boldsymbol{V}$ and unobserved input $\boldsymbol{\theta}_d^u$ is that $\boldsymbol{V}$ is *not* caused by treatment D while we allow $\boldsymbol{\theta}_d^u$ to be determined by treatment. Thus, $\boldsymbol{V}_1 \overset{dist}{=} \boldsymbol{V}_0$, where $\overset{dist}{=}$ means equal in distribution. We discuss the relationship between unobserved variables $\boldsymbol{\theta}_d^u$ and $\boldsymbol{V}$ in presenting our mediation model.

Model (a) of Fig. 1 represents the standard confounding model as a (DAG).[1] In this model, $(Y_1, Y_0) \perp\!\!\!\perp D$ does not hold due to confounding effects of unobserved variables $\boldsymbol{V}$. As a consequence, the observed empirical relationship between output Y and treatment D is not causal and ATE cannot be evaluated by the conditional difference in means between treated and untreated subjects, i.e., $E(Y \mid D = 1) - E(Y \mid D = 0)$. Nevertheless, if $\boldsymbol{V}$ were observed, ATE could be identified from $\int E(Y \mid D = 1, \boldsymbol{V} = v) - E(Y \mid D = 0, \boldsymbol{V} = v) dF_{\boldsymbol{V}}(v)$ as $(Y_1, Y_0) \perp\!\!\!\perp D \mid \boldsymbol{V}$ holds.

The literature on matching (Rosenbaum and Rubin, 1983) solves the problem of confounders by assumption. It postulates that a set of observed preprogram variables, say $\boldsymbol{X}$, spans the space generated by unobserved variables $\boldsymbol{V}$ although it offers no guidance on how to select this set. Thus it assumes that observed preprogram variables $\boldsymbol{X}$ can be found such that $(Y_1, Y_0) \perp\!\!\!\perp D \mid \boldsymbol{X}$ holds. In this case, ATE can be computed by

$$E(Y_1 - Y_0) = \int E(Y \mid D = 1, X = x) - E(Y \mid D = 0, X = x) dF_X(x).$$

For a review of matching assumptions and their limitations, see Heckman and Navarro (2004) and Heckman and Vytlacil (2007).

Randomized controlled trials solve the problem of confounders by design. A standard RCT model for confounders can be represented by the following five variables: (1) an output of interest Y; (2) a treatment indicator D that causes the output of interest and

[1] See Pearl (2009) and Heckman and Pinto (2012) for discussions of causality and Directed Acyclic Graphs.

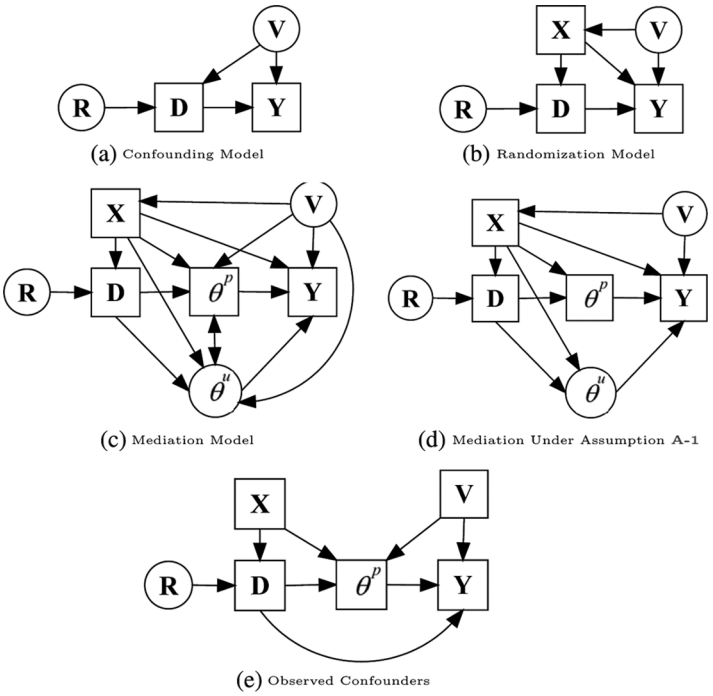

FIGURE 1 Mechanisms of causality for treatment effects. This chart represents five causal models as directed acyclic graphs. Arrows represent causal relationships. Circles represent unobserved variables. Squares represent observed variables. Y is an output of interest. $\boldsymbol{V}$ are unobserved variables. D is the treatment variable. $\boldsymbol{X}$ are preprogram variables. R is the random device used in RCT models to assign treatment status. $\boldsymbol{\theta}^p$ are measured inputs. $\boldsymbol{\theta}^u$ are unmeasured inputs. Both $\boldsymbol{\theta}^p$ and $\boldsymbol{\theta}^u$ play the role of mediation variables. Figure (a) shows a standard confounding model. Figure (b) shows a general randomized trial model. Figure (c) shows a general mediation model where unobserved variables $\boldsymbol{V}$ cause mediation variable $(\boldsymbol{\theta}^p, \boldsymbol{\theta}^u)$. Figure (d) shows the causal relationships of a mediation model that are allowed to exist for Assumption A-1 to hold. Figure (e) shows the mediation model presented in Pearl (2001).

is generated by a random device R and variables $\boldsymbol{X}$ used in the randomization protocol; (3) pre-program variables $\boldsymbol{X}$ used in the randomization protocol; (4) a random device R that assigns treatment status; and (5) an unobserved variable $\boldsymbol{V}$ that causes both $\boldsymbol{X}$ and Y. Model (b) in Fig. 1 represents the RCT model as a directed acyclic graphs (DAG).

In the RCT model, potential confounding effects of unobserved variables $\boldsymbol{V}$ are eliminated by observed variables $\boldsymbol{X}$. ATE can be identified by

$$E(Y_1 - Y_0) = \int E(Y \mid D = 1, X = x) - E(Y \mid D = 0, X = x) dF_X(x).$$

While $(Y_1, Y_0) \perp\!\!\!\perp D \mid \boldsymbol{X}$ holds in both matching and RCT models, it holds by assumption in matching models and by design in RCT models.

We now examine mediation analysis under the assumption that treatment status is generated by a RCT. To this end, we explicitly include measured and unmeasured inputs $(\boldsymbol{\theta}^p, \boldsymbol{\theta}^u)$ to our RCT framework depicted in Model (b) of Fig. 1. Inputs mediate treatment effects, i.e., inputs are caused by D and cause Y. Moreover, we also allow preprogram variables $\boldsymbol{X}$ to cause mediators $\boldsymbol{\theta}^p, \boldsymbol{\theta}^u$. The most general mediation model is described by the following relationships: (R1) mediators $\boldsymbol{\theta}^p, \boldsymbol{\theta}^u$ are caused by unobserved variable $\boldsymbol{V}$; and (R2) measured inputs can cause unmeasured ones and vice-versa. Model (c) of Fig. 1 represents this mediation model for RCT as a DAG.

A production function representation that rationalizes the mediation model is

$$Y_d = f_d(\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_d^u, \boldsymbol{V}, \boldsymbol{X}), d \in \{0, 1\}. \tag{11}$$

Equation (11) differs from Eq. (3) by explicitly introducing preprogram unobserved variables $\boldsymbol{V}$. $Y_{d,\bar{\boldsymbol{\theta}}^p}$ is now defined as

$$Y_{d,\bar{\boldsymbol{\theta}}^p} = f_d(\bar{\boldsymbol{\theta}}_d^p, \boldsymbol{\theta}_d^u, \boldsymbol{V}, \boldsymbol{X}). \tag{12}$$

It is the variation in $\boldsymbol{\theta}_d^u$, $\boldsymbol{V}$, and $\boldsymbol{X}$ that generate randomness in outcome Y_d, fixing $\bar{\boldsymbol{\theta}}^p$.

We gain further insight into Assumption A-1 by examining it in light of the mediation model. The mediation model is constructed under the assumption that treatment status is generated by a RCT. Therefore, Condition (i) of Assumption A-1 holds. However, randomization does not generate Condition (ii) of Assumption A-1. If either $R1$ or $R2$ occurs, measured and unmeasured inputs will not be independent conditioned on observed variables $(D, \boldsymbol{X})$. As a consequence, $Y_{d,\bar{\boldsymbol{\theta}}_d^p} \not\perp\!\!\!\perp \boldsymbol{\theta}_{d'}^p \mid D, \boldsymbol{X}$. Model (d) of Fig. 1 represents a mediation model in which Assumption A-1 holds, but neither $R1$ nor $R2$ occurs.

Condition (ii) is stronger than the conditions invoked in conventional matching analyses. Indeed, if $\boldsymbol{V}$ is assumed to be observed (a matching assumption), then relationship $R1$ reduces to a causal relationship among observed variables. Nevertheless, the matching assumption does not rule out $R2$. Relationship $R2$ would not apply if we adopt the strong assumption that unmeasured inputs have no variation conditional on the treatment.[2] The no-variation assumption assures that measured and unmeasured mediators are statistically independent conditional on D. This model is represented as a DAG in Model (e) of Fig. 1. Pearl (2001) shows why Condition (ii) will not hold for Model (e) of Fig. 1. However, the direct effect (Eq. 5) can be computed by Condition (ii):

$$DE(d) = \iint (E(Y \mid D=1, \boldsymbol{\theta}^p = t, V = v) - E(Y \mid D=0, \boldsymbol{\theta}^p = t, V = v)) dF_V(v) \times \int dF_{\boldsymbol{\theta}^p \mid D=d, \boldsymbol{X}=x, V=v}(t) dF_{\boldsymbol{X}}(x).$$

[2]Pearl (2001) invokes this assumption.

A general solution to the mediation problem is outside the scope of this paper. Instead we use a linear model to investigate how experimental variation coupled with additional econometric exogeneity assumptions can produce a credible mediation analysis for the case where some inputs are unobserved (but may be changed by the experiment) and proxied variables $\boldsymbol{\theta}^p$ are measured with error. Our analysis is based on the production function defined in Eq. (3). We assume that the map between inputs $\boldsymbol{\theta}_d^p, \boldsymbol{\theta}_d^u$ and output Y_d is given by a linear function. We then show how multiple measures on inputs and certain assumptions about the exogeneity of inputs allow us to test for invariance, i.e., whether $f_1(\cdot)$ is equal to $f_0(\cdot)$. Alternatively, invoking invariance we show how to test the hypothesis that *increments* in $\boldsymbol{\theta}_d^p$ are statistically independent of $\boldsymbol{\theta}_d^u$.

3. A LINEAR MODEL FOR MEDITATION ANALYSIS

We focus on examining a linear model for the production function of output in sector d. The benefit of the linear model stems from its parsimony in parameters, which facilitates reliable estimation in small samples. Nonlinear or nonparametric procedures require large samples often not available in RCTs. We write

$$Y_d = \kappa_d + \boldsymbol{\alpha}_d \boldsymbol{\theta}_d + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d, \qquad d \in \{0, 1\}, \tag{13}$$

where κ_d is an intercept, $\boldsymbol{\alpha}_d$ and $\boldsymbol{\beta}_d$ are, respectively, $|\mathcal{J}|$-dimensional and $|\boldsymbol{X}|$-dimensional vectors of parameters, where $|\boldsymbol{Q}|$ denotes the number of elements in $\boldsymbol{Q}$. Preprogram variables $\boldsymbol{X}$ are assumed not to be affected by the treatment. Their effect on Y can be affected by the treatment. $\tilde{\epsilon}_d$ is a zero-mean error term assumed to be independent of regressors $\boldsymbol{\theta}_d$ and $\boldsymbol{X}$.

Technology (13) is compatible with a Cobb–Douglas model using linearity in logs. Thus an alternative to (13) is

$$\log(Y_d) = \kappa_d + \boldsymbol{\alpha}_d \log(\boldsymbol{\theta}_d) + \boldsymbol{\beta}_d \log(\boldsymbol{X}) + \tilde{\epsilon}_d, \qquad \text{or} \tag{14}$$

$$Y_d = \kappa_d + \boldsymbol{\alpha}_d \log(\boldsymbol{\theta}_d) + \boldsymbol{\beta}_d \log(\boldsymbol{X}) + \tilde{\epsilon}_d, \qquad d \in \{0, 1\}. \tag{15}$$

We discuss the estimation of $\boldsymbol{\theta}_d$ in Section 3.1. There, we also adopt a linear specification for the measurement system that links unobserved inputs $\boldsymbol{\theta}$ with measurements M. The Cobb–Douglas specification can be applied to the linear measurement system by adopting a linear-in-logs specification in the same fashion as used in outcome Eqs. (14) and (15).

Analysts of experiments often collect an array of measures of the inputs. However, it is very likely that there are relevant inputs not measured. We decompose the term $\boldsymbol{\alpha}_d \boldsymbol{\theta}_d$ in Eq. (13) into components due to inputs that are measured and inputs that are not:

$$Y_d = \kappa_d + \sum_{j \in \mathcal{J}} \alpha_d^j \theta_d^j + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d$$

$$= \kappa_d + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{inputs on which we} \\ \text{have measurements}}} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{inputs on which we} \\ \text{have no measurements}}} + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d$$

$$= \tau_d + \sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j + \boldsymbol{\beta}_d \boldsymbol{X} + \epsilon_d, \tag{16}$$

where $d \in \{0, 1\}$, $\tau_d = \kappa_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j E(\theta_d^j)$, and ϵ_d is a zero-mean error term defined by $\epsilon_d = \tilde{\epsilon}_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - E(\theta_d^j))$. Any differences in the error terms between treatment and control groups can be attributed to differences in the inputs on which we have no measurements. Without loss of generality, we assume that $\tilde{\epsilon}_1 \overset{dist}{=} \tilde{\epsilon}_0$, where $\overset{dist}{=}$ means equality in distribution. Note that the error term ϵ_d is correlated with the measured inputs if measured inputs are correlated with unmeasured inputs.

We seek to decompose treatment effects into components attributable to changes in the inputs that we can measure. Assuming that changes in unmeasured inputs attributable to the experiment are independent of $\boldsymbol{X}$, treatment effects can be decomposed into components due to changes in inputs $E(\Delta\theta^j)$ and components due to changes in parameters $\Delta\alpha^j (= \alpha_1^j - \alpha_0^j)$:

$$\begin{aligned} E(\Delta Y_d | \boldsymbol{X}) &= E(Y_1 - Y_0 | \boldsymbol{X}) \\ &= (\tau_1 - \tau_0) + E\left(\sum_{j \in \mathcal{J}_p} \left(\alpha_1^j \theta_1^j - \alpha_0^j \theta_0^j \right) \right) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \boldsymbol{X} \\ &= (\tau_1 - \tau_0) \\ &\quad + \sum_{j \in \mathcal{J}_p} \left(\left(\Delta\alpha^j + \alpha_0^j \right) E(\Delta\theta^j) + (\Delta\alpha^j) E\left(\theta_0^j \right) \right) \\ &\quad + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \boldsymbol{X}.^3 \end{aligned} \tag{17}$$

Equation (17) can be simplified if treatment affects inputs, but not the impact of inputs and background variables on outcomes, i.e., $\alpha_1^j = \alpha_0^j$; $j \in \mathcal{J}_p$ and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$.[4] This says that all treatment effects are due to changes in inputs. Under this assumption, the term associated with $\boldsymbol{X}$ drops from the decomposition. Note that under this assumption there still may be a direct effect (Eq. 5) but it arises from experimentally induced shifts in unmeasured inputs.

If measured and unmeasured inputs are independent in the no-treatment outcome equation, $\boldsymbol{\alpha}_0$ can be consistently estimated by standard methods. Under this assumption,

[3] Alternative decompositions are discussed below in Section 6.1.

[4] These are called structural invariance or autonomy assumptions in the econometric literature. See, e.g., Hurwicz (1962). These assumptions do not rule out heterogenous responses to treatment because $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_0$ may vary in the population.

we can test if the experimentally-induced *increments* in unmeasured inputs are independent of the experimentally induced *increments* in measured inputs. This allows us to test a portion of Condition (ii) of Assumption A-1. The intuition for this test is as follows. The inputs for treated participants are the sum of the inputs they would have had if they were assigned to the control group plus the increment due to treatment. If measured and unmeasured input increments are independent, $\boldsymbol{\alpha}_1$ is consistently estimated by standard methods, and we can test $H_0 : plim\ \hat{\boldsymbol{\alpha}}_1 = plim\ \hat{\boldsymbol{\alpha}}_0$, where $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_0)$ are least squares estimators of $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0)$. Notice that even if $\hat{\boldsymbol{\alpha}}_0$ is not consistently estimated, the test of the independence of the increments from the base is generally valid. Assuming the exogeneity of $\boldsymbol{X}$, we can also test if $plim\ \hat{\boldsymbol{\beta}}_1 = plim\ \hat{\boldsymbol{\beta}}_0$.

Note further that if we maintain that measured inputs are independent of unmeasured inputs for both treatment and control groups, we can test the hypothesis of autonomy $H_0 : \boldsymbol{\alpha_1} = \boldsymbol{\alpha_0}$. Thus there are two different ways to use the data from an experiment (a) to test the independence of the increments given that unmeasured inputs are independent of measured inputs or (b) to test $H_0 : \boldsymbol{\alpha_1} = \boldsymbol{\alpha_0}$ maintaining full independence.

Imposing autonomy simplifies the notation. Below we show conditions under which we can test for autonomy. Equation (16) can be expressed as

$$Y_d = \tau_d + \sum_{j_\epsilon \mathcal{J}} \boldsymbol{\alpha}^{\boldsymbol{j}} \theta^j + \boldsymbol{\beta X} + \boldsymbol{\epsilon}_d,\ d \in \{0, 1\}. \tag{18}$$

In this notation, the observed outcome can be written as

$$\begin{aligned} Y &= \underbrace{D\,(\tau_1 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_1^j + \boldsymbol{\beta X} + \boldsymbol{\epsilon}_1)}_{Y_1} + \underbrace{(1 - D)\,(\tau_0 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_0^j + \boldsymbol{\beta X} + \boldsymbol{\epsilon}_0)}_{Y_0} \\ &= \tau_0 + \tau D + \sum_{j \in \mathcal{J}_p} \alpha^j \theta^j + \boldsymbol{\beta X} + \boldsymbol{\epsilon}, \end{aligned} \tag{19}$$

where $\tau = \tau_1 - \tau_0$ is the contribution of unmeasured variables to mean treatment effects, $\boldsymbol{\epsilon} = D\boldsymbol{\epsilon}_1 + (1 - D)\boldsymbol{\epsilon}_0$ is a zero-mean error term, and $\theta^j = D\theta_1^j + (1 - D)\theta_0^j,\ j \in \mathcal{J}_p$ denotes the inputs that we can measure.

If the $\theta_d^j,\ j \in \mathcal{J}_p$ are measured without error and are independent of the error term $\boldsymbol{\epsilon}$, least squares estimators of the parameters of Eq. (19) are unbiased for $\alpha^j,\ j \in \mathcal{J}_p$. If, on the other hand, the unmeasured inputs are correlated with both measured inputs and outputs, least squares estimators of $\alpha^j,\ j \in \mathcal{J}_p$, are biased and capture the effect of changes in the unmeasured inputs as they are projected onto the measured components of $\boldsymbol{\theta}$, in addition to the direct effects of changes in measured components of $\boldsymbol{\theta}$ on Y.

The average treatment effect is

$$E(Y_1 - Y_0) = \underbrace{(\tau_1 - \tau_0)}_{\substack{\text{treatment effect due} \\ \text{to unmeasured inputs}}} + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha^j E(\theta_1^j - \theta_0^j)}_{\substack{\text{treatment effect due} \\ \text{to measured inputs}}}. \tag{20}$$

Input j can explain treatment effects only if it affects outcomes ($\alpha^j \neq 0$) and, on average, is affected by the experiment ($E(\theta_1^j - \theta_0^j) \neq 0$). Using experimental data, it is possible to test both conditions.

Decomposition (20) would be straightforward to identify if the measured variables are independent of the unmeasured variables, and the measurements are accurate. The input term of Eq. (20) is easily constructed by using consistent estimates of the α^j and the effects of treatment on inputs. However, measurements of inputs are often riddled with measurement error. We next address this problem.

3.1. Addressing the Problem of Measurement Error

We assume access to multiple measures on each input. This arises often in many studies related to the technology of human skill formation. For example, there are multiple psychological measures of the same underlying development trait (see, e.g., Cunha and Heckman, 2008, and Cunha et al., 2010). More formally, let the index set for measures associated with factor $j \in \mathcal{J}_p$ be $\mathcal{M}^j$. Denote the measures for factor j by $M^j_{m^j,d}$, where $m^j \in \mathcal{M}^j$, $d \in \{0,1\}$. $\boldsymbol{\theta}_d$ denotes the vector of factors associated with the inputs that can be measured in treatment state d, i.e., $\boldsymbol{\theta}_d = (\theta_d^j : j \in \mathcal{J}_p), d \in \{0,1\}$.

We assume that each input measure is associated with at most one factor. The following equation describes the relationship between the measures associated with factor j and the factor:

$$\text{Measures} : M^j_{m^j,d} = v^j_{m^j} + \varphi^j_{m^j}\theta^j_d + \eta^j_{m^j},\ j \in \mathcal{J}_p,\ m^j \in \mathcal{M}^j. \tag{21}$$

To simplify the notation, we keep the covariates $\boldsymbol{X}$ implicit. Parameters $v^j_{m^j}$ are measure-specific intercepts. Parameters $\varphi^j_{m^j}$ are factor loadings. The $\boldsymbol{\epsilon}_d$ in (18) and $\eta^j_{m^j}$ are mean-zero error terms assumed to be independent of $\boldsymbol{\theta}_d, d \in \{0,1\}$, and of each other. The factor structure is characterized by the following equations:

$$\text{Factor Means} : E[\theta^j_d] = \mu^j_d,\ j \in \mathcal{J}_p \tag{22}$$

$$\text{Factor Covariance} : \text{Var}[\boldsymbol{\theta}_d] = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d},\ d \in \{0,1\}. \tag{23}$$

The assumption that the parameters $v^j_{m^j}$, $\varphi^j_{m^j}$, $\text{Var}(\eta^j_{m^j}) : m^j \in \mathcal{M}^j$, $j \in \mathcal{J}_p$, do not depend on d simplifies the notation, as well as the interpretation of the estimates obtained from our procedure. It implies that the effect of treatment on the measured inputs

operates only through the latent inputs and not the measurement system for those inputs. However, these assumptions are not strictly required. They can be tested by estimating these parameters separately for treatment and control groups and checking if measurement equation factor loadings and measurement equation intercepts differ between treatment and control groups.

4. IDENTIFICATION

Identification of factor models requires normalizations that set the location and scale of the factors (e.g., Anderson and Rubin, 1956). We set the location of each factor by fixing the intercepts of one measure—designated "the first"—to zero, i.e., $\nu_1^j = 0$, $j \in \mathcal{J}_p$. This defines the location of factor j for each counterfactual condition. We set the scale of the factor by fixing the factor loadings of the first measure of each skill to one, i.e., $\varphi_1^j = 1$, $j \in \mathcal{J}_p$. For all measures that are related to a factor (i.e., have a nonzero loading on the factor, $\varphi_{m^j}^j$), the decomposition of treatment effects presented in this paper is invariant to the choice of which measure is designated as the "first measure" for each factor provided that the normalizing measure has a nonzero loading on the input. The decompositions are also invariant to any affine transformations of the measures. Our procedure can be generalized to monotonic nonlinear transformations of the measures.

Identification is established in four steps. First, we identify the means of the factors, μ_d^j. Second, we identify the measurement factor loadings $\varphi_{m^j}^j$, the variances $\mathrm{Var}(\eta_{m^j}^j)$ of the measurement system, and the factor covariance structure $\Sigma_{\boldsymbol{\theta}_d}$. Third, we use the parameters identified from the first and second steps to secure identification of the measurement intercepts $\nu_{m^j}^j$. Finally, we use the parameters identified in the first three steps to identify the factor loadings $\boldsymbol{\alpha}$ and intercept τ_d of the outcome equations. We discuss each of these steps.

1. **Factor Means.** We identify μ_1^j and μ_0^j from the mean of the designated first measure for treatment and control groups: $E(M_{1,d}^j) = \mu_d^j$, $j \in \mathcal{J}_p, d \in \{0,1\}$.
2. **Measurement Loadings.** From the covariance structure of the measurement system, we can identify as follows: (a) the factor loadings of the measurement system $\varphi_{m^j}^j$; (b) the variances of the measurement error terms, $\mathrm{Var}(\eta_{m^j}^j)$; and (c) the factor covariance matrix, $\Sigma_{\boldsymbol{\theta}_d}$. Factors are allowed to be freely correlated. We need at least three measures for each input $j \in \mathcal{J}_p$, all with nonzero factor loadings. The $\varphi_{m^j}^j$ can depend on $d \in \{0,1\}$, and we can identify $\varphi_{m^j,d}^j$. Thus we can test if $H_0 : \varphi_{m^j,1}^j = \varphi_{m^j,0}^j$, $j \in \mathcal{J}_p$, and do not have to impose autonomy on the measurement system.
3. **Measurement Intercepts.** From the means of the measurements, i.e., $E(M_{m^j,d}^j) = \nu_{m^j}^j + \varphi_{m^j}^j \mu_d^j$, we identify $\nu_{m^j}^j$, $m^j \in \mathcal{M}^j \backslash \{1\}$, $j \in \mathcal{J}_p$. Recall that the factor loadings $\varphi_{m^j}^j$ and factor means μ_d^j are identified. Assuming equality of the intercepts ($\nu_{m^j}^j$) between treatment and control groups guarantees that treatment effects on measures, i.e., $E(M_{m^j,1}^j) - E(M_{m^j,0}^j)$, operate solely through treatment effects on factor means,

i.e., $\mu_1^j - \mu_0^j$. However, identification of our decomposition requires intercept equality only for the designated first measure of each factor. We can test $H_0 : v^j_{m^j,1} = v^j_{m^j,0}$ for all $m^j \in \mathscr{M}^j \backslash \{1\}, j \in \mathscr{J}_p$, and hence do not have to impose autonomy on the full measurement system.

4. **Outcome Equation.** Outcome factor loadings in Eq. (18) can be identified using the covariances between outcomes and the designated first measure of each input. We form the covariances of each outcome Y_d with the designated first measure of each input $j \in \mathscr{J}_p$ to obtain $\text{Cov}(Y_d, \boldsymbol{M}_{1,d}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d} \boldsymbol{\alpha}$ where $\boldsymbol{\alpha} = (\alpha^j; j \in \mathscr{J}_p)$. By the previous argument, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}$ is identified. Thus $\boldsymbol{\alpha}$ is identified whenever $\det(\Sigma_{\boldsymbol{\theta}_d}) \neq 0$. We do not have to impose autonomy or structural invariance. Outcome factor loadings $\boldsymbol{\alpha}$ can depend on $d \in \{0, 1\}$, as they can be identified through $\text{Cov}(Y_d, \boldsymbol{M}_{1,d}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d} \boldsymbol{\alpha}_d$ which can be identified separately for treatments and controls. We can test $H_0 : \alpha_1^j = \alpha_0^j,\ j \in \mathscr{J}_p$. Using $E(Y_d)$, we can identify τ_d because all of the other parameters of each outcome equation are identified.

5. ESTIMATION PROCEDURE

We can estimate the model using a simple three stage procedure. First, we estimate the measurement system. Second, from these equations we can estimate the skills for each participant. Third, we estimate the relationship between participant skills and outcomes. Proceeding in this fashion makes identification and estimation transparent.

Step 1. For a given set of dedicated measurements, and choice of the number of factors, we estimate the factor model using measurement system (21)–(23). There are several widely used procedures to determine the number of factors. Examples of these procedures are the scree test (Cattell, 1966), Onatski's criterion (2009), and Horn's (1965) parallel analysis test. In addition, the Guttman–Kaiser rule (Guttman, 1954, and Kaiser, 1960) is well known to overestimate the number of factors (see Zwick and Velicer, 1986, Gorsuch, 2003, and Thompson, 2004). We refer to Heckman et al. (2013) for a detailed discussion of the selection of number of factors.

Step 2. We use the measures and factor loadings estimated in the first step to compute a vector of *factor scores* for each participant i. We form unbiased estimates of the true vector of skills $\boldsymbol{\theta}_i = (\theta_i^j; j \in \mathscr{J}_p)$ for agent i. The factor measure equations contain $\boldsymbol{X}$ which we suppress to simplify the expressions. Notationally, we represent the measurement system for agent i as

$$\underbrace{\boldsymbol{M}_i}_{|\mathscr{M}|\times 1} = \underbrace{\boldsymbol{\varphi}}_{|\mathscr{M}|\times|\mathscr{J}_p|} \underbrace{\boldsymbol{\theta}_i}_{|\mathscr{J}_p|\times 1} + \underbrace{\boldsymbol{\eta}_i}_{|\mathscr{M}|\times 1}, \tag{24}$$

where $\boldsymbol{\varphi}$ represents a matrix of the factor loadings estimated in first step and $\boldsymbol{M}_i$ is the vector of stacked measures for participant i subtracting the intercepts $v^j_{m^j}$ of

Equation (21). The dimension of each element in Eq. (24) is shown beneath it, where $\mathcal{M} = \cup_{j \in \mathcal{J}_p} \mathcal{M}^j$ is the union of all the index sets of the measures. The error term for agent i, $\boldsymbol{\eta}_i$, has zero mean and is independent of the vector of skills $\boldsymbol{\theta}_i$. $\text{Cov}(\boldsymbol{\eta}_i, \boldsymbol{\eta}_i) = \boldsymbol{\Omega}$. The most commonly used estimator of factor scores is based on a linear function of measures: $\boldsymbol{\theta}_{S,i} = \boldsymbol{L}'\boldsymbol{M}_i$. Unbiasedness requires that $\boldsymbol{L}'\boldsymbol{\varphi} = \boldsymbol{I}_{|\mathcal{J}|}$, where $\boldsymbol{I}_{|\mathcal{J}|}$ is a $|\mathcal{J}|$-dimensional identity matrix.[5] To achieve unbiasedness, $\boldsymbol{L}$ must satisfy $\boldsymbol{L}' = (\boldsymbol{\varphi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}'\boldsymbol{\Omega}^{-1}$. The unbiased estimator of the factor is

$$\boldsymbol{\theta}_{S,i} = \boldsymbol{L}'\boldsymbol{M}_i = (\boldsymbol{\varphi}'\boldsymbol{\Omega}^{-1}\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}'\boldsymbol{\Omega}^{-1}\boldsymbol{M}_i.$$

Factor score estimates can be interpreted as the output of a *GLS* estimation procedure where measures are taken as dependent variables and factor loadings are treated as regressors. By the Gauss–Markov theorem, for a known $\boldsymbol{\varphi}$ the proposed estimator is the best linear unbiased estimator of the vector of inputs $\boldsymbol{\theta}_i$.[6]

Step 3. The use of factor scores instead of the true factors to estimate Eq. (18) generates biased estimates of outcome coefficients $\boldsymbol{\alpha}$. Even though estimates of $\boldsymbol{\theta}_i$ are unbiased, there is still a discrepancy between the true and measured $\boldsymbol{\theta}_i$ due to estimation error. To correct for the bias, we propose a bias-correction procedure. Because we estimate the variance of $\boldsymbol{\theta}$ and the variance of the measurement errors in the first step of our procedure, we can eliminate the bias created by the measurement error.

Consider the outcome model for agent i,

$$Y_i = \boldsymbol{\alpha}\boldsymbol{\theta}_i + \boldsymbol{\gamma}\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i, \tag{25}$$

where $(\boldsymbol{\theta}_i, \boldsymbol{Z}_i) \perp\!\!\!\perp \boldsymbol{\epsilon}_i$ and $E(\boldsymbol{\epsilon}_i) = 0$. For brevity of notation, we use $\boldsymbol{Z}_i$ to denote preprogram variables, treatment status indicators, and the intercept term of Eq. (18). From Eq. (24), the factor scores $\boldsymbol{\theta}_{S,i}$ can be written as the inputs $\boldsymbol{\theta}_i$ plus a measurement error $\boldsymbol{V}_i$, that is,

$$\boldsymbol{\theta}_{S,i} = \boldsymbol{\theta}_i + \boldsymbol{V}_i \text{ such that } (\boldsymbol{Z}_i, \boldsymbol{\theta}_i) \perp\!\!\!\perp \boldsymbol{V}_i \text{ and } E(\boldsymbol{V}_i) = 0. \tag{26}$$

Replacing $\boldsymbol{\theta}_i$ with $\boldsymbol{\theta}_{S,i}$ yields $Y_i = \boldsymbol{\alpha}\boldsymbol{\theta}_{S,i} + \boldsymbol{\gamma}\boldsymbol{Z}_i + \boldsymbol{\epsilon}_i - \boldsymbol{\alpha}\boldsymbol{V}_i$. The linear regression estimator of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ is inconsistent:

$$\text{plim}\begin{pmatrix}\hat{\boldsymbol{\alpha}}\\ \hat{\boldsymbol{\gamma}}\end{pmatrix} = \underbrace{\begin{pmatrix}\text{Cov}(\boldsymbol{\theta}_S, \boldsymbol{\theta}_S) & \text{Cov}(\boldsymbol{\theta}_S, \boldsymbol{Z})\\ \text{Cov}(\boldsymbol{Z}, \boldsymbol{\theta}_S) & \text{Cov}(\boldsymbol{Z}, \boldsymbol{Z})\end{pmatrix}^{-1}\begin{pmatrix}\text{Cov}(\boldsymbol{\theta}, \boldsymbol{\theta}) & \text{Cov}(\boldsymbol{\theta}, \boldsymbol{Z})\\ \text{Cov}(\boldsymbol{Z}, \boldsymbol{\theta}) & \text{Cov}(\boldsymbol{Z}, \boldsymbol{Z})\end{pmatrix}}_{A}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\gamma}\end{pmatrix}. \tag{27}$$

[5]The method is due to Bartlett (1937) and is based on the restricted minimization of mean squared error, subject to $\boldsymbol{L}'\boldsymbol{\varphi} = \boldsymbol{I}_{|\mathcal{J}|}$.

[6]Note that the assumption that $\boldsymbol{\varphi}$ is known can be replaced with the assumption that $\boldsymbol{\varphi}$ is consistently estimated and we can use an asymptotic version of the Gauss–Markov theorem replacing "unbiased" with "unbiased in large samples." Standard GMM methods can be applied.

This is the multivariate version of the standard one-variable attenuation bias formula. All covariances in $\boldsymbol{A}$ can be computed directly except for the terms that involve $\boldsymbol{\theta}$. $\text{Cov}(\boldsymbol{\theta}, \boldsymbol{\theta})$ is estimated in step (1). Using Eq. (26), we can compute $\text{Cov}(\boldsymbol{Z}, \boldsymbol{\theta}_S) = \text{Cov}(\boldsymbol{Z}, \boldsymbol{\theta})$. Thus, $\boldsymbol{A}$ is identified. Our bias-correction procedure consists of pre-multiplying the least squares estimators $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ by $\boldsymbol{A}^{-1}$, thus providing consistent estimates of $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$.[7] A one-step maximum likelihood procedure, while less intuitive, directly estimates the parameters without constructing the factors and accounts for measurement error. It is justified in large samples under standard regularity conditions.

6. INVARIANCE TO TRANSFORMATION OF MEASURES

We present some invariance results regarding the decomposition of treatment effects under transformations of the measures used to proxy the inputs. Our analysis is divided into two parts. Section 6.1 examines the invariance of the decomposition for affine transformation of measures under the linear model discussed in the previous section. Section 6.2 relaxes the linearity assumption of Section 6.1 and discusses some generalized results for the case of non-linear monotonic transformations using the analysis of Cunha et al. (2010).

6.1. Invariance to Affine Transformations of Measures

We first establish conditions under which outcome decomposition (20), relating treatment effects to experimentally induced changes in inputs, is invariant to affine transforms of any measure of input for any factor. Decomposition (20) assumes $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0$. We also consider forming decompositions for the more general nonautonomous case where $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_0$. We establish the invariance of the treatment effect due to measured inputs (see Eq. (20)) but not of other terms in the decompositions that arise in the more general case. Throughout, we assume autonomy of the *measurement* system so that intercepts and factor loadings are the same for treatments and controls for all measurement equations. Our analysis can be generalized to deal with that case but at the cost of greater notational complexity.

Before presenting a formal analysis, it is useful to present an intuitive motivation. Let $\widetilde{M}^j_{m^j,d}$ be an affine transformation of the measure $M^j_{m^j,d}$, for some $j \in \mathcal{J}_p$ and $m^j \in \mathcal{M}^j$. Specifically, define $\widetilde{M}^j_{m^j,d}$ by

$$\widetilde{M}^j_{m^j,d} = aM^j_{m^j,d} + b \text{ such that } a \in \mathbb{R}\backslash\{0\}, b \in \mathbb{R}, \text{ and } d \in \{0,1\}, \text{ for all } j \in \mathcal{J}_p. \quad (28)$$

Let $\tilde{\varphi}^j_{m^j}, \tilde{\eta}^j_{m^j}, \tilde{\nu}^j_{m^j}$ be the factor loading, error term, and intercept associated with the transformed measure $\widetilde{M}^j_{m^j,d}, d \in \{0,1\}$. The key condition for the invariance

[7] See Croon (2002) for more details on this bias correction approach.

of decomposition (20) to linear transformations of the different measures is that $\sum_{j\in\mathcal{J}_p}\alpha^j E(\theta_1^j-\theta_0^j)$ be invariant.

We apply the same normalization to the transformed system as we do to the original system. Suppose that the measure transformed is a "first measure" so $m_j=1$. Recall that in the original system, $v_1^j=0$ and $\varphi_1^j=1$. Transformation (28) can be expressed as

$$\widetilde{M}_{1,d}^j=b+a\theta_d^j+a\eta_1^j.$$

Applying the normalization rule to this equation defines factor $\tilde{\theta}_j=b+a\theta_j$, i.e., the scale and the location of the factor are changed, so that in the transformed system the intercept is 0 and the factor loading 1:

$$\widetilde{M}_{1,d}^j=\tilde{\theta}_d^j+\tilde{\eta}_1^j,$$

where $\tilde{\eta}_1^j=a\eta_1^j$ is a rescaled mean zero error term. This transformation propagates through the entire system, where θ_d^j is replaced by $\tilde{\theta}_d^j$.

Notice that in decomposition (20), the induced shift in the mean of the factor is irrelevant. It differences out in the decomposition. The scale of θ^j is affected. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta_d}}$ is transformed to $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}}}$ where

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}}}=\boldsymbol{I}_a\boldsymbol{\Sigma}_{\boldsymbol{\theta_d}}\boldsymbol{I}_a,$$

where $\boldsymbol{I}_a$ is a square diagonal matrix of the same dimension as the number of measured factors and the jth diagonal is a and the other elements are unity. The factor loading for the outcome function for the set of transformed first measures, $\widetilde{\boldsymbol{M}}_{\boldsymbol{1,d}}=\boldsymbol{M}_{1,d}\boldsymbol{I}_a$, is the solution to the system of equations

$$\mathrm{Cov}(Y_d,\widetilde{\boldsymbol{M}}_{1,d})=\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}}}\tilde{\boldsymbol{\alpha}}_d.$$

Thus

$$\begin{aligned}\tilde{\boldsymbol{\alpha}}_{\boldsymbol{d}}&=\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}}}^{-1}\mathrm{Cov}(\boldsymbol{Y}_d,\widetilde{\boldsymbol{M}}_{1,d})\\&=\boldsymbol{I}_a^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta_d}}^{-1}\boldsymbol{I}_a^{-1}\mathrm{Cov}(\boldsymbol{Y}_d,\widetilde{\boldsymbol{M}}_{\boldsymbol{1,d}})\\&=\boldsymbol{I}_a^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta_d}}^{-1}\mathrm{Cov}(\boldsymbol{Y}_d,\boldsymbol{M}_{1,d})\\&=\boldsymbol{I}_a^{-1}\boldsymbol{\alpha_d}.\end{aligned}$$

Since $\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}}=\boldsymbol{I}_a\boldsymbol{\theta_d}$, it follows trivially that decomposition (20), $\boldsymbol{\alpha}_{\boldsymbol{D}}'(\boldsymbol{\theta_1}-\boldsymbol{\theta_0})$, is invariant to transformations.

Suppose next that the transformation is applied to any measure other than a first measure. Invoking the same kind of reasoning, it is evident that $\tilde{\boldsymbol{\theta}}_{\boldsymbol{d}}=\boldsymbol{\theta_d}$ and $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{d}}=\boldsymbol{\alpha_d}$. Thus

the decomposition is invariant. Clearly, however, the intercept of the transformed measure becomes

$$\tilde{v}^j_{m_j} = b + a v^j_{m_j}$$

and the factor loading becomes

$$\tilde{\varphi}^j_{m_j} = \varphi^j_{m_j} a.$$

The preceding analysis assumes that the outcome system is autonomous: $\boldsymbol{\alpha_0} = \boldsymbol{\alpha_1}$, and $\boldsymbol{\beta_0} = \boldsymbol{\beta_1}$. Suppose that $\boldsymbol{\alpha_1} \neq \boldsymbol{\alpha_0}$. To simplify the argument, we continue to assume that $\boldsymbol{\beta_0} = \boldsymbol{\beta_1}$. In this case,

$$E(Y_1 - Y_0) = E(\boldsymbol{\alpha_1}'\boldsymbol{\theta_1} - \boldsymbol{\alpha_0}'\boldsymbol{\theta_0}).$$

In the general case, the decomposition is not unique due to a standard index number problem. Using the notation $\Delta\boldsymbol{\alpha} = \boldsymbol{\alpha_1} - \boldsymbol{\alpha_0}$,

$$\begin{aligned} E(Y_1 - Y_0) &= \underbrace{\boldsymbol{\alpha_0}' E(\boldsymbol{\theta_1} - \boldsymbol{\theta_0})}_{\substack{\text{invariant to affine} \\ \text{transformations of} \\ \text{measures}}} + \underbrace{(\Delta\boldsymbol{\alpha})' E(\boldsymbol{\theta_1})}_{\substack{\text{non invariant to affine} \\ \text{transformations of} \\ \text{measures}}} \\ &= \underbrace{\boldsymbol{\alpha_1}' E(\boldsymbol{\theta_1} - \boldsymbol{\theta_0})}_{\substack{\text{invariant to affine} \\ \text{transformations of} \\ \text{measures}}} - \underbrace{(\Delta\boldsymbol{\alpha}) E(\boldsymbol{\theta_0})}_{\substack{\text{non invariant to affine} \\ \text{transformations of} \\ \text{measures}}} . \end{aligned}$$

For any $\boldsymbol{\alpha}^*$ that is an affine transformation of $(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$,

$$E(Y_1 - Y_0) = (\boldsymbol{\alpha}^*) E(\boldsymbol{\theta_1} - \boldsymbol{\theta_0}) + (\boldsymbol{\alpha_1} - \boldsymbol{\alpha}^*) E(\boldsymbol{\theta_1}) - (\boldsymbol{\alpha_0} - \boldsymbol{\alpha}^*) E(\boldsymbol{\theta_0}).$$

For all three decompositions, the first set of terms associated with the mean change in skills due to treatment is invariant to affine transformations. The proof follows from the preceding reasoning. Any scaling of the factors is offset by the revised scaling of the factor loadings.

Notice, however, that when $\boldsymbol{\alpha_1} \neq \boldsymbol{\alpha_0}$, in constructing decompositions of treatment effects, we acquire terms in the level of the factors. For transformations to the first measure, the change in the *location* is shifted. Even though the scales of $(\Delta\boldsymbol{\alpha})$ and $E(\boldsymbol{\theta_d})$ offset, there is no compensating shift in the location of the factor. Thus the terms associated with the levels of the factor are not, in general invariant to affine transformations of first measures although the decompositions are invariant to monotonic transformations of any non-normalization measures. Obviously the point of evaluation against $E(\boldsymbol{\theta_1} - \boldsymbol{\theta_0})$ is evaluated depends on the choice of $\boldsymbol{\alpha_0}$, $\boldsymbol{\alpha_1}$, and $\boldsymbol{\alpha}^*$ if they differ.

We now formally establish these results. It is enough to consider the transformation of one measure within group j for treatment category d. First, suppose that the

transformation (28) is not applied to the first measure, that is, $m^j \neq 1$. In this case, $E(\theta_1^j - \theta_0^j)$; $j \in \mathcal{J}_p$ are invariant as they are identified through the first measure of each factor which is not changed. We can also show that the α^j, $j \in \mathcal{J}_p$, are invariant. We identify $\boldsymbol{\alpha} = [\alpha^j;\ j \in \mathcal{J}_p]$ through $\mathrm{Cov}(Y_d, \boldsymbol{M}_{1,d}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}\boldsymbol{\alpha}$. Therefore, it suffices to show that covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}$ is invariant under the linear transformation (28). But the covariance between the factors is identified through the first measure of each factor. The variance of factor j under transformation (28) is identified by

$$\begin{aligned}
\frac{\mathrm{Cov}(M_{1,d}^j, \widetilde{M}_{m,d}^j)\mathrm{Cov}(M_{1,d}^j, M_{m',d}^j)}{\mathrm{Cov}(\widetilde{M}_{m,d}^j, M_{m',d}^j)} &= \frac{\mathrm{Cov}(M_{1,d}^j, aM_{m,d}^j)\mathrm{Cov}(M_{1,d}^j, M_{m',d}^j)}{\mathrm{Cov}(aM_{m,d}^j, M_{m',d}^j)} \qquad \text{by (28)} \\
&= \frac{a\mathrm{Cov}(M_{1,d}^j, M_{m,d}^j)\mathrm{Cov}(M_{1,d}^j, M_{m',d}^j)}{a\mathrm{Cov}(M_{m,d}^j, M_{m',d}^j)} \\
&= \frac{\mathrm{Cov}(M_{1,d}^j, M_{m,d}^j)\mathrm{Cov}(M_{1,d}^j, M_{m',d}^j)}{\mathrm{Cov}(M_{m,d}^j, M_{m',d}^j)} \\
&= \mathrm{Var}(\theta_d^j),
\end{aligned}$$

so that the variance is unchanged. Hence $\boldsymbol{\alpha_d}$ is unchanged.

Now suppose that transformation (28) is applied to the first measure, $m^j = 1$. In this case, the new variance of factor j is given by

$$\begin{aligned}
\frac{\mathrm{Cov}(\widetilde{M}_{1,d}^j, M_{m,d}^j)\mathrm{Cov}(\widetilde{M}_{1,d}^j, M_{m',d}^j)}{\mathrm{Cov}(M_{m,d}^j, M_{m',d}^j)} &= \frac{a\mathrm{Cov}(M_{1,d}^j, M_{m,d}^j)a\mathrm{Cov}(M_{1,d}^j, M_{m',d}^j)}{\mathrm{Cov}(M_{m,d}^j, M_{m',d}^j)} \\
&= a^2\mathrm{Var}(\theta_d^j). \qquad (29)
\end{aligned}$$

The new covariance between factors j and j' is given by:

$$\begin{aligned}
\mathrm{Cov}(\widetilde{M}_{1,d}^j, M_{1,d}^{j'}) &= a\mathrm{Cov}(M_{1,d}^j, M_{1,d}^{j'}) \\
&= a\mathrm{Cov}(\theta_d^j, \theta_d^{j'}). \qquad (30)
\end{aligned}$$

Let $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_d}$ be the new factor covariance matrix obtained under transformation (28). According to Eqs. (29) and (30), $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_d} = \boldsymbol{I}_a\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}\boldsymbol{I}_a$, where, as before, $\boldsymbol{I}_a$ is a square diagonal matrix whose jth diagonal element is a and has ones for the remaining diagonal elements. By the same type of reasoning, we have that the covariance matrix $\mathrm{Cov}(Y_d, \boldsymbol{M}_{1,d})$ computed under the transformation is given by $\mathrm{Cov}(Y_d, \widetilde{\boldsymbol{M}}_{1,d}) = \boldsymbol{I}_a\mathrm{Cov}(Y_d, \boldsymbol{M}_{1,d})$. Let $\tilde{\boldsymbol{\alpha}}$ be the outcome factor loadings under transformation (28). Thus,

$$\boldsymbol{I}_a\mathrm{Cov}(Y_d, \boldsymbol{M}_{1,d}) = \mathrm{Cov}(Y_d, \widetilde{\boldsymbol{M}}_{1,d}) = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}_d}\tilde{\boldsymbol{\alpha}} = \boldsymbol{I}_a\boldsymbol{\Sigma}_{\boldsymbol{\theta}_d}\boldsymbol{I}_a\tilde{\boldsymbol{\alpha}}, \qquad (31)$$

and therefore, $\tilde{\boldsymbol{\alpha}} = \boldsymbol{I}_a^{-1}\boldsymbol{\alpha}$. In other words, transformation (28) only modifies the jth factor loading, which is given by $\tilde{\alpha}^j = \frac{\alpha^j}{a}$.

Let the difference in factor means between treatment groups be $\Delta^{j'} = E(\theta_1^{j'} - \theta_0^{j'}), j' \in \mathcal{J}_p$, and let $\widetilde{\Delta}^{j'}$ be the difference under transformation (28). Transformation (28) only modifies the jth difference in means which is given by $\widetilde{\Delta}^j = a\Delta^j$ and thereby $\tilde{\alpha}^j\widetilde{\Delta}^j = \alpha^j\Delta^j$. Thus $\tilde{\alpha}^{j'}\widetilde{\Delta}^{j'} = \alpha^{j'}\Delta^{j'} = \alpha^{j'}E(\theta_1^{j'} - \theta_0^{j'})$ for all $j' \in \mathcal{J}_p$, as claimed. It is straightforward to establish that if $\boldsymbol{\alpha_1} \neq \boldsymbol{\alpha_0}$, the decomposition is, in general, not invariant to affine transformations, although the term associated with $E(\boldsymbol{\theta_1} - \boldsymbol{\theta_0})$ is.

6.2. A Sketch of More General Invariance Results

We next briefly consider a more general framework. We draw on the analysis of Cunha et al. (2010) to extend the discussion of the preceding subsection to a nonlinear nonparametric setting. We present two basic results: (1) outcome decomposition terms that are locally linear in $\boldsymbol{\theta}$ are invariant to monotonic transformations of $\boldsymbol{\theta}$; and (2) terms associated with shifts in the technology due to the experimental manipulation are not. In this section, we allow inputs to be measured with error but assume that unmeasured inputs are independent of the proxied ones. We focus only on invariance results and only sketch the main ideas.

Here we sketch the main results. Following the previous notation, we use D for the binary treatment status indicator, $D = 1$ for treated, and $D = 0$ for control. We denote $Y_d;\ d \in \{0, 1\}$ to denote the output Y when treatment D is fixed at value d. In the fashion, $\boldsymbol{\theta}_d;\ d \in \{0, 1\}$ denotes the input $\boldsymbol{\theta}$ when treatment D is fixed at value d. For sake of simplicity, let the production function be given by $f : \text{supp}(\boldsymbol{\theta}) \rightarrow \text{supp}(Y)$ where supp means support. Thus $Y_d = f(\boldsymbol{\theta}_d); d \in \{0, 1\}$.

We analyze both the invariant and noninvariant case. We relax the invariance assumption for the production function by indexing it by treatment status. We use $f_d : \text{supp}(\boldsymbol{\theta}) \rightarrow \text{supp}(Y)$ to denote the production function that governs the data generating process associated with treatment status $D = d$.

In this notation, the average treatment effect is given by

$$E(Y_1 - Y_0) = E(f_1(\boldsymbol{\theta}_1) - f_0(\boldsymbol{\theta}_0)). \tag{32}$$

Equation (32) repeats the discussion in Section 2 that there are two sources of treatment effects exist: (1) treatment might shift the map between θ and the outcomes from f_0 to f_1 (i.e., it might violate invariance); and (2) treatment might also change the inputs from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}_1$.

Assume the existence of multiple measures of $\boldsymbol{\theta}$ that are generated through an unknown function $M : \text{supp}(\boldsymbol{\theta}) \rightarrow \text{supp}(M)$ that is monotonic in $\boldsymbol{\theta}$. Then, under conditions specified in Cunha et al. (2010), the marginal distributions of $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_0}$ can be nonparametrically

identified (although not necessarily the joint distribution of $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_0}$). We develop the scalar case.

Theorem T-1. The Scalar Case. *Let the production function be a uniformly differentiable scalar function* $f_d : \text{supp}(\theta) \to \text{supp}(Y)$; $d \in \{0, 1, \}$. *If the production function is autonomous, i.e.,* $f_1(t) = f_0(t)\ \forall\ t \in \text{supp}(\theta)$, *then the effect attributable to changes in* θ *is invariant to monotonic transformations* M *of* θ.

Proof. Without loss of generality, write the input for treated in terms of the input for untreated plus the difference across inputs. Thus $\theta_1 = \theta_0 + \Delta$. Now, under structural invariance,

$$Y_1 - Y_0 = f(M(\theta_1)) - f(M(\theta_0)) = f(M(\theta_0 + \Delta)) - f(M(\theta_0)).$$

From uniform differentiability of M and f, we have that

$$\lim_{\Delta \to 0} \frac{Y_1 - Y_0}{\Delta} = \frac{\partial f}{\partial M} \frac{\partial M}{\partial \theta}.$$

Thus the infinitesimal contribution of a change in input to output can be decomposed as

$$d(Y_1 - Y_0) = \frac{\partial f}{\partial M} \frac{\partial M}{\partial \theta} d\theta.$$

If we use θ as the argument of the function, under conditions specified in Cunha et al. (2010), nonparametric regression identifies $\frac{\partial f}{\partial M} \frac{\partial M}{\partial \theta}$. If we use $M(\theta)$ as the argument, nonparametric regression identifies $\frac{\partial f}{\partial M}$, but the increment to input is now $\frac{\partial M}{\partial \theta} d\theta$. The combined terms for the output decomposition remain the same in either case. Thus the decomposition is invariant to monotonic transformations M of inputs θ. Extension to the vector case is straightforward.

Suppose that we relax autonomy. For sake of simplicity, take the scalar case, and let the input for the treated input be written as $\theta_1 = \theta_0 + \Delta$. In this case, we can write the total change in output induced by treatment as

$$\begin{aligned} Y_1 - Y_0 &= f_1(M(\theta_1)) - f_0(M(\theta_0)) \\ &= f_1(M(\theta_0 + \Delta)) - f_0(M(\theta_0)) \\ &= (f_1(M(\theta_0 + \Delta)) - f_1(M(\theta_0))) + (f_1(M(\theta_0)) - f_0(M(\theta_0))). \end{aligned}$$

If we rework the rationale of the proof for Theorem T-1 and apply the intermediate value theorem, we obtain the following expression:

$$Y_1 - Y_0 = \underbrace{\frac{\partial f_1}{\partial M}\frac{\partial M}{\partial \theta}\bigg|_{\theta=\theta_0^*}}_{\text{Invariant}} \Delta\theta + \underbrace{f_1(M(\theta_0)) - f_0(M(\theta_0))}_{\text{Non-Invariant}}, \tag{33}$$

where θ_0^* is an intermediate value in the interval $(\theta_0, \theta_0 + \Delta)$. The first term is invariant for the same reasons stated in Theorem T-1 which concerns the autonomous case. Namely, the change in $\frac{\partial f}{\partial M}$ offsets the change in $\frac{\partial M}{\partial \theta}$.

The source of non-invariance of the second term in Eq. (33) is attributed to the shift in production function f_0 to f_1 due to treatment. This shift implies that the output evaluation will differ when evaluated at the same input points θ_0. Under structural invariance or autonomy, $f_1(\cdot) = f_0(\cdot)$, and regardless of the transformation M, we have that $f_1(M(\theta_0)) = f_0(M(\theta_0))$ and, therefore, the second term of Eq. (33) vanishes.

7. SUMMARY AND CONCLUSIONS

Randomization identifies treatment effects for outputs and measured inputs. If there are unmeasured inputs that are statistically dependent on measured inputs, unaided experiments do not identify the causal effects of measured inputs on outputs.

This paper reviews the recent statistical mediation literature that attempts to identify the causal effect of measured changes in inputs on treatment effects. We relate it to conventional approaches in the econometric literature. We show that the statistical mediation literature achieves its goals under implausibly strong assumptions. For a linear model, we relax these assumptions maintaining exogeneity assumptions that can be partially relaxed if the analyst has access to experimental data. Linearity gives major simplifying benefits even in the case where $\boldsymbol{\theta}_d^p$ is independent of $\boldsymbol{\theta}_d^u$, where the point of evaluation of mean effects does not depends on the distribution of $\boldsymbol{\theta}_d^u$. Extension of this analysis to the nonlinear case is a task left for future work.

We also present results for the case where there is measurement error in the proxied inputs, a case not considered in the statistical literature. When the analyst has multiple measurements on the mismeasured variables, it is sometimes possible to circumvent this problem. We establish invariance to the choice of monotonic transformations of the input measures for both linear and nonlinear technologies.

ACKNOWLEDGMENTS

The views expressed in this paper are those of the authors and not necessarily those of the funders or persons named here. We thank the editor and an anonymous referee for helpful comments.

FUNDING

This research was supported in part by the American Bar Foundation, the JB & MK Pritzker Family Foundation, Susan Thompson Buffett Foundation, NICHD R37HD065072, R01HD54702, a grant to the Becker Friedman Institute for Research and Economics from the Institute for New Economic Thinking (INET), and an anonymous funder. We acknowledge the support of a European Research Council grant hosted by University College Dublin, DEVHEALTH 269874.

REFERENCES

Anderson, T. W., Rubin, H. (1956). Statistical inference in factor analysis. In: Neyman, J. (ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 5. Berkeley, CA: University of California Press, pp. 111–150.

Baron, R. M., Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51:1173–1182.

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology* 28:97–104.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* 1(2):245–276.

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In: Marcoulides, G. A., Moustaki, I. eds., *Latent Variable and Latent Structure Models*. New Jersey.: Lawrence Erlbaum Associates, Inc., pp. 195–223.

Cunha, F. (2012). Eliciting maternal beliefs about the technology of skill formation. Presented at, Family Inequality Network: Family Economics and Human Capital in the Family, November 16, 2012.

Cunha, F., Heckman, J. J. (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources* 43(4):738–782.

Cunha, F., Heckman, J. J., Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica* 78(3):883–931.

Gorsuch, R. L. (2003). Factor analysis. In: Weiner, I. B., Freedheim, D. K., Schinka, J. A., Velicer, W. F. eds., *Handbook of Psychology: Research Methods in Psychology*. Vol. 2, Chapter 6. Hoboken, NJ: John Wiely & Sons, Inc., pp. 143–164.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika* 19(2):149–161.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* 11(1):1–12.

Heckman, J., Holland, M., Oey, T., Olds, D., Pinto, R., Rosales, M. (2012). A reanalysis of the nurse family partnership program: The memphis randomized control trial. Unpublished manuscript, University of Chicago.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., Yavitz A. Q. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics* 1(1):1–46.

Heckman, J. J., Navarro, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics* 86(1):30–57.

Heckman, J. J., Pinto, R. (2012). Causal analysis after Haavelmo: Definitions and a unified analysis of identification. Unpublished manuscript, University of Chicago, Chicago, IL, USA.

Heckman, J. J., Pinto, R., Savelyev, P. A. (2013). Understanding the mechanisms through which an Influential early childhood program boosted adult outcomes. *American Economic Review* 103(6):1–35.

Heckman, J. J., Vytlacil, E. J. (2007). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In: Heckman, J., Leamer, E., eds., *Handbook of Econometrics*. Vol. 6B, Chapter 71. Amsterdam: Elsevier, pp. 4875–5143.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30(2):179–185.

Hurwicz, L. (1962). On the structural form of interdependent systems. In: Nagel, E., Suppes, P., Tarski, A., eds.), *Logic, Methodology and Philosophy of Science*. Stanford, CA: Stanford University Press, pp. 232–239.

Imai, K., Keele, L., Tingley, D., Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review* 105(4):765–789.

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

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 20(1):141–151.

Kaiser, H. F. (1961). A note on Guttman's lower bound for the number of common factors. *British Journal of Statistical Psychology* 14(1):1–2.

Klein, L. R., Goldberger, A. S. (1955). *An Econometric Model of the United States, 1929–1952*. Amsterdam: North-Holland Publishing Company.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann Publishers Inc., pp. 411–420.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, (2nd ed.), New York: Cambridge University Press.

Pearl, J. (2011). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. Forthcoming in *Causality: Statistical Perspectives and Applications*.

Robin, J.-M. (2003). *Comments on Structural Equations, Treatment Effects and Econometric Policy Evaluation*, by James J. Heckman and Edward Vytlacil. Presented at the Sorbonne, Paris.

Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Theil, H. (1958). *Economic Forecasts and Policy*. Amsterdam: North-Holland Publishing Company.

Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*. Washington, D.C.: American Psychological Association.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* 20:557–585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics* 5(3):161–215.

Zwick, W. R., Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin* 99(3):432–442.