Poor economics and its missing mechanisms: The case for causal mediation

Sunil Mitra Kumar* Ragupathy Venkatachalam†

Abstract

A key aim of studying development is to understand the factors that shape socioeconomic progress and explain inequalities. In empirical work, the predominant focus has been on posing these questions in the language of causal inference: how one or more variables effect an outcome of interest, with the estimation of Average Treatment Effects (ATE) becoming prioritized as the key objective. The 'credibility revolution' and the emphasis on randomized controlled trials in research on development has cemented this dominance, because randomization is well-suited to estimating the ATE. This paper argues that this dual dominance – ATE as main question of interest, and experiment as preferred method – is narrow and restrictive. We propose causal mediation frameworks as an alternative, which are routinely used in disciplines including epidemiology, psychology, sociology and political science where causal mechanisms are an equally important focus. We introduce key concepts and definitions of path-specific effects, and discuss identification and estimation approaches. We illustrate applications for development, and demonstrate how causal mediation brings the focus back to contextual knowledge and combine it with empirical rigour.

Keywords: Causal mediation, causal mechanisms, development economics

1 Introduction

Among the key aims of the study of development across disciplines is to understand the factors that shape socioeconomic progress and explain underlying, often persistent variations across groups and countries. Such factors can operate and be defined at macro level (e.g. nations, institutions, structural constraints) and micro level (e.g. individuals and households). In practice, interest in this larger query is often broken down into

*King's India Institute and Department of International Development, King's College London
†Institute of Management Studies, Goldsmiths, University of London, London.

smaller sub-questions at different levels to make theoretical and empirical investigations tractable. These involve focusing on isolating a smaller subset of variables, developing a theoretical framework to explain observed empirical patterns, potential determinants, and where relevant offer policy insights.

In development economics, the predominant focus has been on posing these questions in the language of causal inference: how one or more variables effect an outcome of interest. This interest in causal inference has shaped the field in distinct ways, with the estimation of Average Treatment Effects (ATE) becoming prioritized as the key objective, be it using experimental or observational data. For example, the effects of a job-training programme for the unemployed on future employment levels, or of a remedial teaching intervention on learning scores among school pupils and so on.

Experimental methods in particular have enjoyed disproportionate attention in the past two decades and there is a sizable literature on their advantages (e.g. Banerjee and Duflo, 2011; Haynes et al., 2012) and key limitations, both methodological and ethical (e.g. Deaton and Cartwright, 2018; Teele, 2014). The ATE focuses attention on the causal effect of a single variable on a single outcome. Randomisation directly enables this form of causal inference, while observational data can also be used to infer these effects using appropriate statistical adjustment.

However in our view, this dual dominance – of the ATE as main question of interest, and experiment as preferred method – represents a narrow and restrictive form of causal inference. The limitation is twofold. First, a narrow focus on *what* works has consequently led to neglecting causal *mechanisms* or why it works. The processes through which various factors together shape developmental outcomes are seldom the object of inquiry, or at best indirectly so. This is in contrast to disciplines like epidemiology, psychology, sociology and political science where causal mechanisms are an important focus. Second, a failure to fully exploit the power of causal inference frameworks to glean insights into such mechanisms from observational data has further narrowed the scope of scholarship.

In contrast, understanding development more holistically necessitates moving beyond measuring ATEs alone and the related disproportionate focus on experimental methods. Instead, to broaden the existing scope to account for causal mechanisms and thereby the myriad ways in which context, structure and processes interact to shape developmental outcomes. In order to understand mechanisms, we need to identify and disentangle the pathways through which causal effects are manifested. This involves studying intermediate variables – mediators – and drawing upon contextual background knowledge to specify the causal structures involved. This also requires bringing the emphasis back to practitioner expertise and contextual or domain knowledge, which otherwise have a tendency to be relatively deprioritised.

In this paper we build this argument and examine the role of causal mediation frame-

works to achieve this. Section 2 provides the motivation behind causal mediation and introduces key concepts: graphical representation of causal structures in the form of Directed Acyclic Graphs (DAGs), and causal estimands of interest such as direct and indirect effects. Section 3 offers a formal discussion of identification and estimation strategies for causal effects and extensions within this framework. Section 4 presents four distinct examples of how causal mediation can be put to work in the context of development, drawing on work from multiple contexts. Section 5 concludes by discussing some challenges and extensions offered by this framework.

2 Mediation: Why and What

Starting with the premise that our interest is in uncovering mechanisms behind development outcomes, what is needed is a conceptual framework that a) enables us to specify these mechanisms with the requisite level of granularity b) where feasible, allows estimating various causal effects. Field experiments in development economics have largely focused on policy or treatment evaluation, measuring the total effects of a given intervention on the outcome of interest, often through the ATE or similar estimand. This focuses on the question of *whether* or *what* works, and the magnitude of the effect. While this is a legitimate question and can be important for programme evaluation, it is not the only question of interest. It is vital to understand *why* and *how* these effects ultimately manifest, which calls for investigating the causal pathways through which actual or hypothetical interventions are translated into outcomes in certain populations.

Mediators are intermediate variables that transmit causal effects from actual or hypothetical interventions to outcomes. Causal mediation analysis aims to specify these channels of transmission, and disentangle and estimate the direct and indirect effects that arise from these (Imai et al., 2010,?; Pearl, 2001; VanderWeele, 2015). A common and effective way to represent causal structures is through the use of Directed Acyclic Graphs or DAGs, which can flexibly accommodate several simultaneous channels through which causal effects manifest (see Chen et al., 2018; Robins, 2003). While the traditional approach to causal mediation analysis proposed by Baron and Kenny (1986) specifies linear models for the outcome and mediator and assigns causal interpretation to model coefficients, the modern approach integrates graphical methods with the potential outcomes framework of Rubin (1974) and the pathway analysis pioneered by Sewall Wright (1934).[1]. This approach has been largely developed and applied within disciplines including sociology, psychology, epidemiology and political science, but less so in economics and

[1]See Pearl (2001), VanderWeele (2015) on the use of the potential outcomes framework for causal mediation, and see Celli (2022) on the historic origins of this approach in the context of economics where, albeit, this method is not widely used.

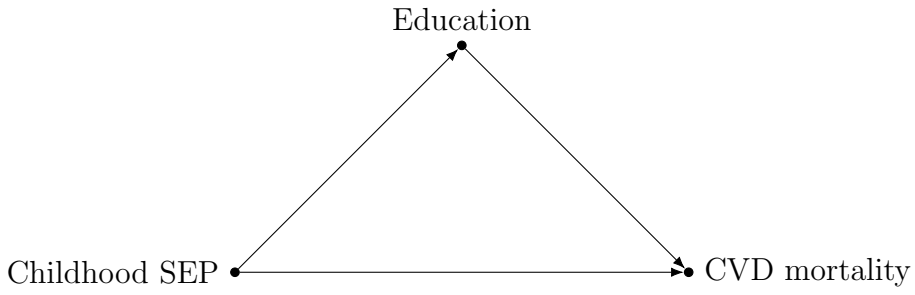

Figure 1: Causal structure for effect of childhood SEP on CVD mortality

econometrics.[2] The key insights yielded by this framework lie in its ability to define and measure path-specific effects.

Below, we describe different aspects of the causal mediation framework: specifying causal structures, and intermediate variables using DAGs, defining estimands of interest, conditions under which these path-specific effects can be identified, and related estimation strategies. We note that this framework is general enough to accommodate non and semi-parametric approaches.

We can motivate this framework through the following example. Childhood socioeconomic position (SEP) is believed to influence adult cardiovascular mortality. Hossin et al. (2021) posit that this relationship is mediated by social (adult education and SEP) and behavioural factors (smoking, alcohol drinking, physical inactivity, poor diet and body mass index). We discuss this example in greater detail in §1, but for now let us consider a simplified version with a single mediator, viz. education. Thus, childhood SEP effects the education level of the individual (as an adult), which in turn effects their cardiovascular (CVD) mortality (within the study period). Childhood SEP also influences cardiovascular mortality through causal pathways other than education. To simplify further, suppose that childhood SEP takes on only two values, viz. high (H) and low (L).

The average treatment effect (ATE) in this context is the difference in expected counterfactual CVD mortality rates between those with high and low childhood SEP, that is, $E[\text{CVD}(\text{SEP} = H)] - E[\text{CVD}(\text{SEP} = L)]$. The corresponding randomised experiment – conceptual in this case – is for childhood SEP to be manipulated. Estimating the ATE using observational data requires adjusting for all possible endogenous variables, for instance location or race since these can effect both childhood SEP as well as adult mortality. In the terminology of mediation, the ATE is known as the Total Causal Effect or TCE, which naturally suggests the existence of other constituent mediated and non-mediated effects. Given figure 1 and the role of education therein, the indirect effect of childhood SEP via education is the change in mortality due to the change in educa-

[2] As Imbens (2020, p.1145) points out, causal mediation has not gained wide acceptance within economics, but it has the potential to help explain causal pathways and perhaps deserves more attention in the discipline.

tion that results from SEP switching from H to L. The corresponding direct effect of childhood SEP is the resulting change in mortality while holding education fixed at some given level.

It is worth noting one crucial distinction between the direct effect and the ATE. Both require (conceptually or otherwise) manipulating SEP. But while the direct effect requires holding education fixed so as to isolate the effect of SEP alone, the ATE requires *ignoring* education as a covariate. Were educated to be adjusted for (e.g. by adding it as a control variable in an OLS regression), the resulting ATE estimate would be biased since it will subsume both the true ATE as well as the indirect effect via education, which in this case presumably magnifies the effects of childhood SEP. More generally, it is a well-known principle in the causal inference literature that merely adjusting for intermediate variables (here education) leads to biased estimates of the ATE (Gelman et al., 2020; Acharya et al., 2016). Randomization – in this case conceptually manipulating childhood SES – can enable estimating the ATE. However, because education is itself effected by childhood SES, randomisation does not in general enable estimating direct or indirect effects via education, a point to which we return below.

3 Path specific effects, their identification, and estimation

Let us start with a simple causal structure of the type given in figure 2 with a single treatment variable (A), single mediator (M), and single outcome (Y). This can be extended to accommodate multiple mediators and more complex causal relationships. Several causal effects of interest can be defined. In what follows, M_a refers to the value M would take when $A = a$, and likewise $M_{a'}$ refers to the value M would take when $A = a'$. Similarly, Y_{am} refers to the value Y would take when $A = a$ and $M = m$, and $Y_{a'M_{a''}}$ refers to the value Y would take when $A = a'$ and $M = M(a'')$.[3]

The overall treatment effect or *Total Causal Effect* (TCE) refers to the average change in Y corresponding to a change in A from $A = a$ to $A = a'$, encompassing all possible causal pathways – those involving mediators as well as those that operate independently of mediators. Formally

$$\text{TCE} = E[Y_{aM_a} - Y_{a'M_{a'}}] \tag{1}$$

Direct effects are those which sidestep any and all mediators, and reflect changes in the outcome Y due to a change in treatment A while holding mediator M fixed at some level

[3]We provide these definitions for the case of continuous Y and M but they are easily extended to the case when either or both of Y and M are categorical. See for instance

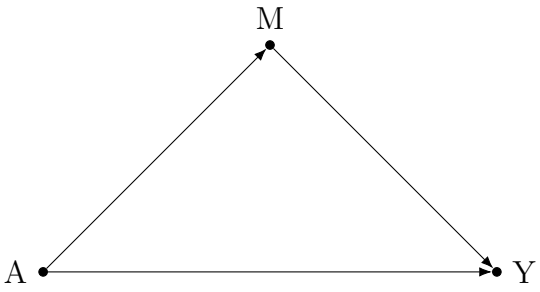

Figure 2: Causal structure for effect of treatment A on outcome Y with mediator M

so as to rule out any changes in Y due to a subsequent change in M. The *Controlled Direct Effect* (CDE) refers to the average change in Y due to a change in A while holding M fixed – controlled – at some given level m for all units. We can also define *natural* direct effects, where the word natural conveys that we are measuring the average change in Y due to a change in A when holding M, for each unit, at the level it would naturally attain at the reference level of treatment. Depending on the reference levels considered, there are two corresponding natural direct effects. In both cases, we vary the treatment from $A = a$ to $A = a'$. The Pure Natural Direct Effect (PNDE) is defined by holding M at M_a for each unit, while the Total Natural Direct Effect (TNDE) is defined by holding M is held at M'_a for each unit. To be clear, in the case of CDE, M is held at some $M = m$ for all units, whereas for the PNDE and TNDE, holding M at the level it would naturally attain at the corresponding levels of treatment implies that those M will likely vary across units. Formally

$$\text{CDE} = E[Y_{am} - Y_{a'm}] \tag{2}$$

$$\text{PNDE} = E[Y_{a'M_a} - Y_{aM_a}] \tag{3}$$

$$\text{TNDE} = E[Y_{a'M_{a'}} - Y_{aM_{a'}}] \tag{4}$$

For instance, in the example set out above, the CDE could be variously defined depending on the level at which education is conceptually held fixed, while manipulating SEP. For the TNDE (say), we let each individual's education attain the level it *naturally* would if their SEP=H, and while holding education fixed at this level, estimate the change in average CVD mortality when instead SEP=L.

Indirect effects are those which operate via mediators, or in the simplest case as in figure 2, a single mediator. These sidestep any changes in Y due to the direct effect of A, thereby focusing exclusively on the 'knock-on' effects wherein A effects M and M effects Y. The conceptual manipulation in this case is to vary the mediator from the level it naturally attains when $A = a$ to that attained when $A = a'$ while holding the treatment itself fixed. The level at which the treatment is held gives rise to definitions

corresponding to those above. The Pure Natural Indirect Effect (PNIE) is defined by holding the treatment at $A = a$, while the Total Natural Indirect Effect (TNIE) is defined by holding the treatment at $A = a'$. In both cases, the causal quantity of interest is the average difference in Y resulting exclusively from the change in $M = M_a$ to $M = M'_a$ for each unit.[4,5] Formally

$$\text{PNIE} = E[Y_{aM_{a'}} - Y_{aM_a}] \tag{5}$$

$$\text{TNIE} = E[Y_{a'M_{a'}} - Y_{a'M_a}] \tag{6}$$

For the example above, the PNIE of childhood SEP is the expected change in CVD mortality that arises, holding SEP fixed at SEP $= L$, from the change in education levels from $E_{\text{SEP}=L}$ to $E_{\text{SEP}=H}$. This is arguably a quantity of interest to, say, a policymaker who wants to understand the likely change in mortality due to an intervention that targets education, changing education levels from $E_{\text{SEP}=L}$ to $E_{\text{SEP}=H}$. That is, this quantity compares CVD mortality in two *natural* states of the world: when everyone attained the education they would with low SEP vs that attained with high SEP, even if the SEP itself stays fixed (in this case at SEP $= L$). Finally, we note that the total causal effect (eq. 1) of a change from $A = a$ to $A = a'$ can be decomposed as the sum of the PNDE and TNIE, or of the PNIE and TNDE (see Pearl, 2001).

3.1 Identification

In essence, identifying the various causal effects defined above requires being able to identify the constituent causal relationships, viz. treatment-mediator, mediator-outcome conditional on treatment, and treatment-outcome conditional on mediator. These can be expressed formally as follows, under the additional overall assumption that the data is generated from a Non-Parametric Structural Equation Model (Pearl, 2009, 2014). [6]

A1: **Consistency**: (i) Consistency of A on M and consistency of $\{A, M\}$ on Y. This assumption requires that actual and potential values coincide for all relevant variables.

A2: **No unobserved mediator-treatment confounding**: There exists a set of variables W_1 such that the effect of treatment on the mediator can be identified by conditioning on W_1. That is, $M_a \perp\!\!\!\perp A \mid W_1$.

[4]There is generally no indirect effect analogous to the CDE because this would involve holding M at two essentially arbitrary levels.

[5]Both types of natural direct and indirect effects will be equal (viz. PNDE=TNDE and PNIE=TNIE) if there are no interaction terms AM involving treatment and mediator in the expression that determines the outcome.

[6]See VanderWeele (Ch.2, 2015) and Pearl (2014) for details.

A3: **No unobserved mediator-outcome confounding**: There exists a set of variables W_2 such that for each value of the treatment $a, a' \in A$, $Y_{am} \perp\!\!\!\perp \mathrm{M} \mid A = a', \{W_1, W_2\}$. In other words, holding treatment A fixed, the causal effect of the mediator on the outcome can be identified.[7]

A4: **No unobserved treatment-outcome confounding**: There exists a set of variables W_3 such that $Y_{aM} \perp\!\!\!\perp A \mid \{W_2, W_3\}$. That is, holding M fixed, the effect of treatment on outcome can be identified by conditioning on $\{W_2, W_3\}$.

A significant proportion of the literature uses to a closely related version of these assumptions, usually referred to as sequential conditional independence (See Imai et al. (2010) and Imai et al. (2010)).

Randomisation of the treatment guarantees A2. Additionally, randomising of the mediator too guarantees A3. But even if both applications of randomisation are feasible, identifying the natural effects defined above also requires A4, for which randomisation alone is insufficient because mediators have to be held at the levels they would *naturally* attain at a certain level of treatment. In general therefore, randomisation is not sufficient for identifying natural direct and indirect effects (see Pearl, 2014, p.460).

3.2 Estimation strategies

Multiple estimation strategies have been proposed in the literature. The traditional, fully parametric approach consists of specifying linear models for the mediator and outcome, and interpreting coefficient estimates in terms of specific causal relationships. As MacKinnon et al. (2020) explain, it is possible to define the direct and indirect effects defined above as functions of coefficients from these models. That is, to obtain equivalent point estimates to specific potential-outcome contrasts, with standard errors obtained via bootstrapping. However, the analytical expressions for the required estimands will need to be re-calculated every time the model specification is changed, and providing analytical expressions for complex or non-linear models can be challenging.

Instead, the two main approaches used in the literature focus on directly estimating contrasts of average potential outcomes rather than deriving analytical expressions. Under the first approach, the outcome is flexibly modelled as a function of treatment, mediators, relevant interaction terms between these, and the covariates needed to satisfy assumptions A2-A4, and each mediator is likewise modelled as a function of treatment, other mediators where relevant including any interaction terms, and other required covariates. Linear or non-linear models can be specified in this step, with interaction and higher-order terms included as appropriate. Next, corresponding to required levels of

[7] Crucially, this assumption also rules out any confounders of the mediator-outcome relationships themselves affected by the treatment, a point to which we return below.

the treatment, predicted values for each mediator across the sample are obtained using the model estimates and covariate values for each unit. Finally, the same step is repeated to obtain predicted values of the outcome using the model estimates by holding the treatment at required levels and replacing the mediator(s) with the predicted values obtained in the previous step. The empirical averages of these predictions provide the point estimates for each expectation term in the definitions above.

Under this approach, the sampling variability of the conditional counterfactual distributions of mediators and outcome are accounted for through Monte-Carlo simulation. This can be done in two ways: by taking draws from the sampling distribution of the model predictions (e.g. Vansteelandt and Daniel, 2017; De Stavola et al., 2015), or from the sampling distribution of the model parameters (e.g. Imai et al., 2010), and both methods can be combined with bootstrapping. A second approach to estimating direct and indirect effects is through inverse probability weighting as suggested by Huber (2014) who demonstrates this approach in the context of a job training programme. This approach directly estimates the sample analogues to eqs. (3)-(6) using two sets of propensity scores for the treatment, conditional on all relevant confounders, and additionally the mediator. Standard errors are again obtained through bootstrapping.

3.3 Multiple mediators and observed confounders

The definitions provided above are for the case of a single mediator. These can be generalised to more complex causal structures involving multiple mediators, however the associated identification requirements can become significantly more arduous depending on the causal interconnections amongst the mediators. A generalised version of this is presented in figure 3, where not only are there pathways from treatment to mediator to outcome, but also potential causal pathways amongst the mediators themselves. A full treatment of this matter is beyond the scope of this paper, however, a brief outline is as follows. The simplest case is when all mediators can be treated as a single bloc.[8] That is, if the researcher wishes to estimate various causal effects for mediators as a whole, and not for individual mediators. Provided assumptions A1-A4 introduced above can be met for each mediator individually and also as as a set, it is straightforward to estimate the total causal effect, and natural direct and indirect effects – the latter treating mediators as a group. The steps outlined in the estimation procedure above are repeated for each mediator, and the required values of Y are simulated by incorporating these accordingly.

Treating mediators other than as a group poses some challenges to identifying certain indirect effects. As a simple example, suppose there are only two mediators M_1 and M_2, and that the researcher wants to estimate the natural indirect effect via M_1, that is, a quantity such as $E[Y_{aM_{1a}M_{2a}} - Y_{aM_{1a'}M_{2a}}]$ where the notation now accounts for two

[8]See VanderWeele and Vansteelandt (2014) for a detailed explanation of this and related cases.

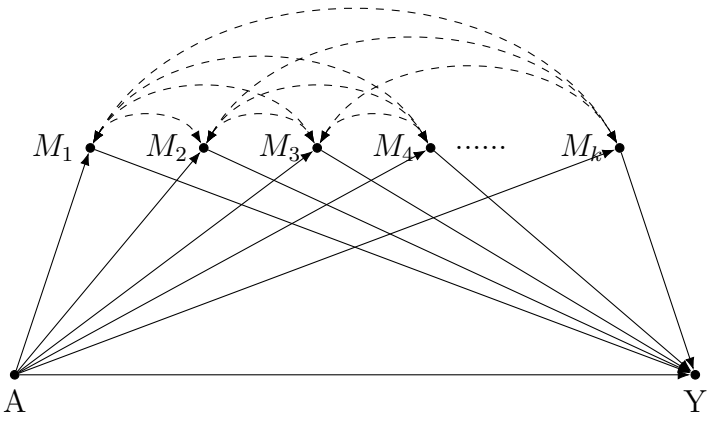

Figure 3: Generalised framework with multiple mediators

Note: Solid arrows represent known causal relationships, while dotted bidirectional arrows represent potential causal relationships

mediators. The first term in this expression is straightforward, as it entails holding the treatment at $A = a$ and both mediators at the level they would attain when $A = a$. The second term involves holding the treatment unchanged, letting M_2 attain the value it would at $A = a$, however holding M_1 at the value it would attain when $A = a'$. Identification is still straightforward provided there is no causal arrow from M_2 to M_1. However, if M_2 effects M_1, obtaining the counterfactual $M_{1a'}$ requires setting $A = a'$ *and* $M_2 = M_{2a'}$, even though $Y_{aM_{1a'}M_{2a}}$ requires $M_2 = M_{2a}$, where the latter requires setting $A = a$. For this reason, $M_{1a'}$ is termed a 'cross-world' counterfactual which cannot be estimated from the data without assuming further structure for the data generating process. Daniel et al. (2015) propose a simple specification for the causal parent M_2 which uses a single parameter, and enables the researcher to probe the sensitivity of any estimates to the value of this parameter.

A final challenge relates to assumption A3 above, that was required in order to identify the mediator-outcome relationship. This assumption rules out any unobserved mediator-outcome confounders, however in general, we additionally need to rule out any *observed* mediator-outcome confounders too if they are themselves effected by the treatment. The reason is that an observed confounder of this sort would be analogous to the M_2 discussed above, viz. a mediator that causally effects a second mediator. De Stavola et al. (2015) provide a detailed treatment of this problem, and offer an implementation of two alternative, weaker sets of assumptions proposed by Robins and Greenland (1992) and Petersen et al. (2006) respectively by into simple constraints on regression coefficients in a structural equation model. VanderWeele et al. (2014) offer a different set of potential solutions: a) to treat all mediators as a block; b) to focus on alternative path-specific effects which also yield a decomposition of the TCE, but are distinct from natural direct

and indirect effects; c) assuming randomisation of the mediator is possible, to estimate a randomised-interventional form of direct and indirect effects where, within each strata of the treatment, the mediator is held at a level randomly chosen from the distribution of the mediator for that strata.

4 Applications in development

Clearly, the concept of causal mediation is not new, and there exists a large literature on identification and estimation, including several methods by which quantities of interest can be estimated.

However, the extent to which mediation as a conceptual framework is used varies widely across disciplines and areas, and we would argue that the study of development is a field where it has found almost no application. This is in stark contrast to the nature of most developmental processes. Howsoever generally that term is defined, causal explanations in social and economic development require delineating multiple interlinked phenomena. Mediation can be useful in broadening the set of conceptual and empirical tools available to the researcher. We now illustrate this argument with four examples.

4.1 Female labour force participation

Labour force participation, particularly for females, is a key indicator of economic development because it touches upon crucial questions related to education, returns thereto, and structure of the economy. The expectation is that female labour force participation (FLFP) will eventually rise with rising income and education, although it might initially decrease before the eventual increase, giving rise to a U-shaped relationship (Klasen, 2019). However, this hypothesis might play out differently depending on historical, cultural and economic context (Klasen et al., 2021; Jayachandran, 2021).

In its simplest form, FLPF should rise due to at least two factors: the demand for labour, and rising productivity due to increases in education. However, at least two factors might counter this. The cultural context might not support the idea of women working outside the household. Second, the nature of marital matching on education might give rise to situations where higher-educated females have higher-educated husbands, resulting in a lower marginal contribution of female earnings to the household. Clearly there are significant factors beyond these as well such as the gender wage gap, demand for higher-educated labour, and regional variation in both.

Conventional approaches to modelling the causal relationship between education and FLFP assign a causal interpretation to the correlation between them, adjusting for factors such as location, regional labour market indicators, household characteristics, and – for cultural contexts where this is relevant – the husband's level of education, earnings, and

other characteristics. However, this approach fails to yield causal estimates, because most of the characteristics adjusted for are not confounders; they are in fact mediators. If we conceptualise labour force participation (FLFP) as the final outcome and the causal variable of interest to be level of education (E), E effects LFP *via* intermediate outcomes: characteristics of the husband following marital matching and the decision of where to live (location and regional labour market conditions). The variables listed above are therefore mediators, which are themselves effected by E, and go on to effect LFP. There might of course also be confounders which need to be added to the model, but these will be in addition to mediators.

Using a causal mediation framework here affords two benefits. It yields unbiased estimates of the total causal effect of education on FLFP, but going beyond this, also enables the researcher to estimate key indirect effects of interest, such as how the nature of marital matching supports FLFP (indirect effect via husband's education and other characteristics), how post-marriage location choices shape FLPF, and how these vary according to categories such as religion or caste. This in turn yields insights that might not otherwise be evident. For example, how shifts over time in the social mores which govern marital matching might change the education-FLFP relationship. The identification requirements are not always easy to fulfill, because as section 3.1 formalises, this requires observing all potential confounders of *each* causal relationship. For example, in contexts where most marriages are arranged, identifying the causal relationship between education and the husband's level of education requires adjusting for the economic and social status of the woman's immediate family – parents – and ideally wider family as well, since they too might contribute to choosing the marital match.

Kumar and Kao (2022) demonstrate this approach in the context of FLFP in India. They focus on the simplest case using only a single mediator, viz. the husband's level of education. They find that positive assortative marital matching – women with higher levels of education marry men with higher levels of education, in a context where most marriages are arranged – depresses FLFP because FLFP falls with rising levels of husband's education. The latter might be due to a combination of cultural factors including caste, and the lack of suitable employment opportunities for educated women (Das, 2006; Klasen and Pieters, 2015). However they do not model location as a mediator in its own right. Doing so would help explain some of the variation due to differences in the availability of employment opportunities, even if not the absolute levels. Similarly, undertaking this same analysis for subsamples by caste group would yield further insight into the role of cultural factors.

4.2 Discrimination

Discrimination poses important challenges for individuals and groups to achieve outcomes commensurate with their capabilities, accentuating inequalities and thereby impeding development in societies. Discrimination often manifests in the form of differences arising due to group identity such as race, gender, caste or other immutable characteristics. However, unequal outcomes across groups may arise for a variety of reasons and therefore not always be directly attributable to discrimination alone. A key aim of studying discrimination is therefore to distinguish it from other mechanisms by which group-belonging potentially gives rise to inequalities. Situations involving discrimination can be readily expressed using the language of causal inference and potential outcomes. Often, the question we seek to answer is the following. "Would the outcome(s) be different had the individual been a member of a different group, with everything else remaining the same?". In other words, we can define, identify and estimate the extent of discrimination in terms of counterfactual contrasts of potential outcomes. Although the potential outcomes framework is widely used within economics to study discrimination, causal mediation can enrich this analysis further.

Various forms of discrimination have been identified in the literature. It might arise due to biased decision-makers who favour certain groups over others. It might also arise in situations with limited, where decision-makers cannot observe a certain relevant attribute (e.g. trustworthiness in a hiring situation) and instead decide to use an individuals' group-belonging (e.g. race or caste) as a proxy. These two distinct forms of discrimination are usually referred to as taste-based and statistical, respectively, but it is clear that they operate through different channels. Distinguishing these channels – causal pathways – is important to understand as well as devise appropriate remedial measures, and it is here that mediation offers an enabling framework.[9]

To illustrate, let us consider a hiring situation where an employer makes decisions based on a candidate's qualifications, trustworthiness, and race. Race can trigger prejudice and leading to taste discrimination, or, as a proxy for beliefs about some unobservable relevant characteristic such as trustworthiness, lead to statistical discrimination. Beliefs about trustworthiness is a mediator here, and the corresponding causal structure is shown in figure 5. In this example, other relevant but observable characteristics that also vary by race such as qualifications, prior experience, and behaviour at interview etc. would be included in the analysis as other mediators as shown in the figure, one or more of which might also potentially effect (or be effected by) beliefs about trustworthiness. The solid black line shows the causal pathway corresponding to taste discrimination, while the dotted pathway is statistical discrimination. Kumar and Venkatachalam (2021) provide a

[9]These are not the only forms of discrimination. For instance, institutions can cause discrimination, where the rules or algorithms by which decisions are made can systematically favour certain groups.

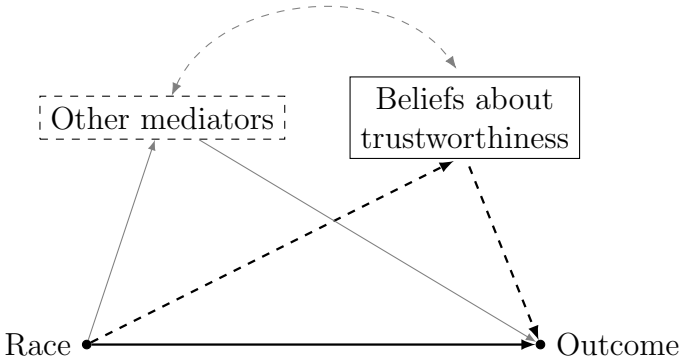

Figure 5: Causal structure for effect of treatment (race) on employment outcome with trustworthiness acting as the mediator.

comprehensive treatment of this problem and a demonstration by way of examining racial discrimination in stop-and-frisk actions by the New York Police Department. Taste discrimination is the Natural Direct Effect of group-belonging (Pearl, 2001), and statistical discrimination as the Natural Indirect Effect which operates via beliefs.[10]

Experimental methods can also be used to study discrimination, but they are some significant limitations in such contexts. First, the idea of randomising race (or caste, gender) poses a practical challenge. Although it can be partially addressed by instead randomising *perceptions* of race (e.g. via candidates' names on a CV as in Bertrand and Mullainathan (2004)), however, challenges still remain. In order to estimate taste discrimination, we need to hold fixed both qualifications as well beliefs about trustworthiness, and the latter presents a challenge, because beliefs need to be held at the level they would naturally attain, for each individual, corresponding to race being set to Black and White. On the other hand, the ideal experiment for measuring statistical discrimination must isolate the effect of beliefs while ensuring race-based prejudice is held fixed. This requires manipulating information about trustworthiness across belief-levels corresponding to, say, Blacks and Whites, while holding perceived race fixed at Black. The practical unfeasibility of both strategies is a special case of the general problem described by Pearl (2014) wherein randomisation is not usually sufficient to recover natural direct and indirect effects. In contrast, causal mediation offers a unified framework for studying taste and statistical discrimination, which can use observational or a mix of observational and experimental data to estimate discrimination.

[10]Note that in the hiring situation discussed here, beliefs about trustworthiness may or may not be directly observable to the researcher. In cases when they are not directly observable, Kumar and Venkatachalam (2021) suggest a simulation-based approach to undertake a bounds-based analysis.

4.3 Political science

Political science is an outlier in the social sciences in that causal mediation is frequently used to study questions about political preferences and decisions through a mixture of experimental and non-experimental approaches. Several of these questions have clear relevance for topics in development. The examples discussed below span causal processes at individual and macro-levels, illustrating the benefits afforded by causal mediation and clear links to questions in development.

Brader et al. (2008) study the effects of media framing on individuals' attitudes and beliefs. They show how attitudes towards immigration are shaped by exposure to news stories about the costs of immigration, and how this varies according to the race of immigrants featured in these stories. They use an experimental design where racial profiles in the news stories were carefully randomised to ensure no other immigrant-features differed significantly. They find that white individuals react more strongly to news stories featuring Latino immigrants and not for European immigrants, and suggesting that these effects might manifest by triggering certain emotions, show that anxiety is the key mediator whose levels rise in response to these news stories. Imai et al. (2011) revisit this analysis, use graphical methods to exemplify the identification assumptions and various causal estimands of interest. Specifically, since treatment (race in the news story) is randomised, the total causal effect on immigration attitudes will be identified without further assumptions. However anxiety is a mediator and is not randomised, so that identifying the indirect effect via anxiety requires ruling out any confounders of the anxiety-attitude relationship.[11]

Through a cross-country comparison, Bormann et al. (2019) examine the effects of formal ethnic power-sharing institutions on ethnic conflict. They posit power-sharing behaviour as the key intermediate outcome and thus mediator, and show that institutions do not have a direct effect on the likelihood of peace, but this increases only if actual power-sharing behaviour – the mediator – changes as well. In other words, the total causal effect can be explained primarily due to the indirect effect. They go on distinguish between two types of conflict: that between members of any power-sharing coalitions, and that between coalition members and non-members. With larger coalitions, the likelihood of infighting rises while that of the second type of conflict decreases. Positing this setup as a question of causal mediation reflects a crucial attribute and contextual insight, namely that power-sharing behaviour might be effected by, but is distinct from and not coincident with power-sharing institutions, and that both go on to shape eventual conflict outcomes. Without this, simply *adjusting* for behaviour changes in a model regressing conflict outcomes on institutional characteristics would not only yield biased estimates,

[11] See Imai et al. (2013) on using sensitivity analyses to probe the role of this and other assumptions in Brader et al. (2008).

but crucially, obscure the insight that changes in institutions manifest through behaviour.

4.4 Policy and program evaluation

Programme and policy evaluation have been a predominant focus of research efforts in development economics, particularly through randomised control trials. A wide array of questions ranging from the effectiveness of macro policies such as development aid, to micro interventions such as availability of microfinance, hiring extra teachers, and use of fertilizers etc. have been investigated in the literature using experimental (RCTs, quasi-experiments, natural experiments) as well as observational data. Within this approach, ATEs have been the prime parameter of interest, even as some studies endeavour to go a step further to measure heterogeneous treatment effects.

The focus on ATEs alone is subject to what is now a well-developed critique. As Deaton (2010, p.424) argues, 'RCT-based evaluation of projects, without guidance from an understanding of underlying mechanisms, is unlikely to lead to scientific progress in the understanding of economic development'. Even within a programme evaluation context, there is need for reliable knowledge not just on *what* works, but mechanisms that shed light on the *why* question and identifying the contexts in which projects may be likely to work. Deaton's critique of Urquiola and Verhoogen (2009) is case in point for the perils of ignoring mechanisms. Causal mediation methods can help ameliorate some of these issues, and we highlight two cases to illustrate this point.

Huber et al. (2017) analyse the effectiveness of caseworkers in enabling unemployed workers to find employment in Switzerland using a rich dataset that links jobseeker and casework data. Behncke et al. (2010) examine how the level of co-operativeness of caseworkers effect the probabilities of employment of their (unemployed) clients. They find that less co-operative caseworkers are better at finding employment for their clients. However, focusing solely on average treatment effects masks the mechanisms that are at play. Huber et al. (2017) go a step further, and decompose the positive ATE associated with less co-operative caseworkers into direct and indirect effects. The indirect effect stems from the assignment to active labour market programmes, while all other causal channels such as the threat of sanctions or the pressure to accept jobs constitute the direct effect. They find that the indirect effect is fact negligible, suggesting that the effectiveness of 'uncooperative' caseworkers stems from aspects of their behaviour in counselling, and not from their effectiveness at assigning clients to suitable labour market programmes. Understanding these mechanisms yields insights concerning how the factors which give rise to this apparent effectiveness might in fact be detrimental in the long run.

The second example involves the effects of the Perry Preschool Programme, an educational intervention during early childhood for disadvantaged children in the USA, on a variety of later-life adult outcomes such as criminal behaviour, health behaviours, and

employment. A large literature studying this intervention, which consistently finds positive treatment effects. Heckman et al. (2013) extend this finding to delve into the causal mechanism. Although they do not explicitly use the language of causal mediation, in essence they focus on the role of two mediators – cognitive and personality skills. They find that it is changes in personality skills due to the intervention – this particular indirect effect – which play a major role in explaining positive overall treatment effects, and not changes in cognitive skills viz. IQ.

5 Discussion

In this paper we have introduced causal mediation frameworks and argued why they are relevant and useful for studying topics in development. We have supported this argument by discussing various applications, and demonstrated how causal questions of several types can benefit from being 'thickened' in this way. This entails moving beyond simple average treatment effect-type queries, and instead positing a causal mechanism of varying complexity, and estimating constituent causal effects of interest. However, doing so is not without challenges.

Each of the examples discussed above begins with a causal mechanism specified a priori. This theorised mechanism allows the researcher to draw a DAG, designate treatment, outcome and mediating variables and propose their causal linkages. Doing so requires domain knowledge, drawing on intuition, observation and field experience, as well as prior empirical and theoretical work. The resulting propositions might therefore be naturally subject to disagreement and some ambiguity. Contrast this with a simple average-treatment-effect type of causal query. Here too prior knowledge and insight is required to pose a question that is relevant and useful, but in essence this concerns only two variables and not the mechanism linking them. Instead, the spirit of causal mediation *necessitates* explicit codification of the causal structure a priori. Once specified, questions of identification and estimation of the parameters of interest are investigated within that structure.

A second challenge is posed by the potential presence of unobserved confounders. Here, unlike estimating the ATE, randomisation can help but cannot fully address the problem if the aim is to estimate natural direct effects. This is a well-known problem in the literature, with several proposals offering sensitivity analysis-based approaches to quantify the magnitude of bias (see Huber, 2020, for a comprehensive overview). In the simpler version of this problem, the confounder is not affected by the treatment. Standard sensitivity approaches involving assumptions on the correlation structure of the confounder and variables of interest (any two of treatment, mediator, outcome) can be employed to obtain bounds on the estimands (e.g. Imai et al., 2010). The more chal-

lenging version is when the treatment effects the unobserved confounder. Tackling this requires stronger assumptions specifying the conditional correlations between treatment, confounder and variables of interest (VanderWeele, 2010).

A third, related challenge concerns the number of mediators that can be included and the associated complexity this introduces for estimation. This has two components. The first is simply the multiplicity of mediator-wise, path-specific effects. For instance, as Daniel et al. (2015) illustrate, n=2 causally ordered mediators give rise to 4!=24 different path specific decompositions of the Total Causal Effect. A closely related challenge is posed by multiple mediators with causal interrelationships. Identifying path-specific effects for individual mediators is not always possible, and sensitivity analyses in response to partial identification might need to contend with several degrees of freedom. Such analyses typically introduce one or more parameters that are systematically varied to estimate sensitivity. This can include parameters to account for either potential unobserved confounders, or with multiple mediators, parameters such as the cross-world correlation parameter proposed by Daniel et al. (2015) discussed in §3.3. As the number of sensitivity parameters increases, the combinations of their values requiring investigation increases exponentially, posing challenges not only of computation, but also how to effectively summarise the results.

Nevertheless, we believe that causal mediation offers a rich conceptual framework with several potential applications in development, some examples of which we have presented above. While we have focused on the basic causal mediation set-up, this can be extended to incorporate other ideas and methods relevant to applications in development. For instance, intersectionality is the study of inequalities and more broadly the differences in experience that arise from the intersection of categories such as sex, race and caste. For instance, the labour market experience of individuals who are female as well as lower-caste might be very different from that predicted by simply conjoining the experiences of females and of lower-castes. These sorts of joint disparities lie at the heart of conceptualising intersectionality. What adds to the complexity is that frequently outcomes of interest (e.g. employment) are determined after an intermediate outcome (e.g. education) which itself displays intersectional disparities. Policy-relevant questions can then be posed, such as the extent to which increasing education levels would lessen the intersectional disparities in employment faced by lower-caste females. This fits naturally within a causal mediation framework as a query about the Controlled Direct Effect of a two-dimensional treatment (sex, caste) where the mediator (education) is being set to some policy-specified level. There is a growing literature on many of these aspects, including how intersectionality can be thought of as a causal phenomenon that assigns counterfactual interpretations to the various categories of belonging (Bright et al., 2016; Jackson et al., 2016; Jackson, 2017).

A second extension arises from bringing mediation analysis to panel data settings. The methods and illustrations discussed in this paper have been proposed for the context of cross-sectional data, albeit bearing in mind the chronological ordering of treatment, mediators and outcome. This can be extended to panel settings where at least one of (treatment, mediators, outcome) are longitudinal. There are multiple complexities to deal with. First, how to define path-specific effects in a panel context, because the values at which (say) the treatment is to be held is a single-dimensional vector for the cross-sectional setting, but multidimensional for panels. Additional complexity can arise if for instance the treatment and mediator are both longitudinal, and there are sequential causal relationships of the sort where treatment at T=0 effects the mediator at T=0 which in turn effects treatment at T=1 and so on. Second, how to extend and adapt identification and estimation strategies for panel settings. VanderWeele and Tchetgen Tchetgen (2017) provide a detailed treatment of these questions as part of a nascent but growing literature.

In conclusion, we believe that there is a need to broaden both methods and scope of enquiry within development economics. Causal mediation analysis provides one way of doing so. It preserves and extends the aims of the 'credibility revolution' while addressing some of the leading criticisms related to, what is in our view, the over-emphasis on experimental approaches. It also has potential to incorporate modern developments in machine learning, and thereby improving estimation and computation strategies and allowing for greater complexity. Most importantly, causal mediation helps enrich knowledge of development processes – and not simply outcomes – by bringing the attention back to mechanisms.

References

Acharya, A., M. Blackwell, and M. Sen (2016, August). Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects. *American Political Science Review 110*(3), 512–529.

Banerjee, A. V. and E. Duflo (2011). *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. London: Penguin.

Baron, R. M. and D. A. Kenny (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology 51*(6), 1173.

Behncke, S., M. Frölich, and M. Lechner (2010). Unemployed and their caseworkers: Should they be friends or foes? *Journal of the Royal Statistical Society: Series A (Statistics in Society) 173*(1), 67–92.

Bertrand, M. and S. Mullainathan (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review 94*(4), 991–1013.

Bormann, N.-C., L.-E. Cederman, S. Gates, B. A. T. Graham, S. Hug, K. W. Strøm, and J. Wucherpfennig (2019). Power Sharing: Institutions, Behavior, and Peace. *American Journal of Political Science 63*(1), 84–100.

Brader, T., N. A. Valentino, and E. Suhay (2008). What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat. *American Journal of Political Science 52*(4), 959–978.

Bright, L. K., D. Malinsky, and M. Thompson (2016, January). Causally Interpreting Intersectionality Theory. *Philosophy of Science 83*(1), 60–81.

Celli, V. (2022). Causal mediation analysis in economics: Objectives, assumptions, models. *Journal of Economic Surveys 36*(1), 214–234.

Chen, B., J. Pearl, and R. Kline (2018). Graphical tools for linear path models. *Psychometrika 4*(5).

Daniel, R. M., B. L. D. Stavola, S. N. Cousens, and S. Vansteelandt (2015). Causal mediation analysis with multiple mediators. *Biometrics 71*(1), 1–14.

Das, M. B. (2006, May). Do Traditional Axes of Exclusion Affect Labor Market Outcomes in India?

De Stavola, B. L., R. M. Daniel, G. B. Ploubidis, and N. Micali (2015, January). Mediation Analysis With Intermediate Confounding: Structural Equation Modeling Viewed Through the Causal Inference Lens. *American Journal of Epidemiology 181*(1), 64–80.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*, 424–455.

Deaton, A. and N. Cartwright (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine 210*, 2–21.

Gelman, A., J. Hill, and A. Vehtari (2020). *Regression and Other Stories*. Cambridge University Press.

Haynes, L., O. Service, B. Goldacre, and D. Torgenson (2012). *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials*. London: UK Cabinet Office and Behavioural Insights Team.

Heckman, J., R. Pinto, and P. Savelyev (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 2052–86.

Hossin, M. Z., I. Koupil, and D. Falkstedt (2021, July). Early life socioeconomic position and mortality from cardiovascular diseases: An application of causal mediation analysis in the Stockholm Public Health Cohort. *BMJ Open 9*(6), e026258.

Huber, M. (2014). Identifying Causal Mechanisms (primarily) Based on Inverse Probability Weighting. *Journal of Applied Econometrics 29*(6), 920–943.

Huber, M. (2020). Mediation Analysis. In K. F. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics*, pp. 1–38. Cham: Springer International Publishing.

Huber, M., M. Lechner, and G. Mellace (2017). Why Do Tougher Caseworkers Increase Employment? The Role of Program Assignment as a Causal Mechanism. *The Review of Economics and Statistics 99*(1), 180–183.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological Methods 15*(4), 309–334.

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

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

Imbens, G. W. (2020, March). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *arXiv:1907.07271 [stat]*.

Jackson, J. W. (2017, July). Explaining intersectionality through description, counterfactual thinking, and mediation analysis. *Social Psychiatry and Psychiatric Epidemiology 52*(7), 785–793.

Jackson, J. W., D. R. Williams, and T. J. VanderWeele (2016, October). Disparities at the intersection of marginalized groups. *Social Psychiatry and Psychiatric Epidemiology 51*(10), 1349–1359.

Jayachandran, S. (2021). Social norms as a barrier to women's employment in developing countries. *IMF Economic Review*, 1–20.

Klasen, S. (2019). What explains uneven female labor force participation levels and trends in developing countries? *The World Bank Research Observer 34*(2), 161–197.

Klasen, S., T. T. N. Le, J. Pieters, and M. Santos Silva (2021). What drives female labour force participation? Comparable micro-level evidence from eight developing and emerging economies. *The Journal of Development Studies 57*(3), 417–442.

Klasen, S. and J. Pieters (2015). What explains the stagnation of female labor force participation in urban India? *The World Bank Economic Review 29*(3), 449–478.

Kumar, S. M. and Y.-F. Kao (2022). Counterfactual thinking and causal mediation: An application to female labour force participation in India. In R. Venkatachalam (Ed.), *Artificial Intelligence, Learning and Computation in Economics and Finance*. Springer Nature.

Kumar, S. M. and R. Venkatachalam (2021). Causal inference for discrimination: An analysis of NYPD stop-and-frisk data.

MacKinnon, D. P., M. J. Valente, and O. Gonzalez (2020, February). The Correspondence between Causal and Traditional Mediation Analysis: The Link is the Mediator by Treatment Interaction. *Prevention science : the official journal of the Society for Prevention Research 21*(2), 147–157.

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

Pearl, J. (2009). *Causality* (Second ed.). New York: Cambridge University Press.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods 19*(4), 459–481.

Petersen, M. L., S. E. Sinisi, and M. J. van der Laan (2006). Estimation of direct causal effects. *Epidemiology*, 276–284.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. *Oxford Statistical Science Series*, 70–82.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688–701.

Teele, D. L. (Ed.) (2014). *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*. New Haven & London: Yale University Press.

Urquiola, M. and E. Verhoogen (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review 99*(1), 179–215.

VanderWeele, T. and S. Vansteelandt (2014, January). Mediation Analysis with Multiple Mediators. *Epidemiologic Methods 2*(1).

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology (Cambridge, Mass.) 21*(4), 540.

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press.

VanderWeele, T. J. and E. J. Tchetgen Tchetgen (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(3), 917–938.

VanderWeele, T. J., S. Vansteelandt, and J. M. Robins (2014, March). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology (Cambridge, Mass.) 25*(2), 300–306.

Vansteelandt, S. and R. M. Daniel (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology 28*(2), 258–265.

Wright, S. (1934). The method of path coefficients. *The annals of mathematical statistics 5*(3), 161–215.