ISSN 2385-2755
DiSSE Working papers
[online]

SAPIENZA
UNIVERSITÀ DI ROMA

WORKING PAPERS SERIES
DIPARTIMENTO DI SCIENZE SOCIALI ED ECONOMICHE

Causal Mediation Analysis in Economics: objectives, assumptions, models

Viviana Celli

N. 12/2019

SAPIENZA - UNIVERSITY OF ROME
P.le Aldo Moro n.5 – 00185 Roma T(+39) 0649910563
CF80209930587 – P.IVA 02133771002

Causal Mediation Analysis in Economics: objectives, assumptions, models

Viviana Celli*

Abstract

The aim of mediation analysis is to identify and evaluate the mechanisms through which a treatment affects an outcome. The goal is to disentangle the total treatment effect into two components: the indirect effect that operates through one or more intermediate variables, called mediators, and the direct effect that captures the other mechanisms. This paper reviews the methodological advancements in causal mediation literature in economics, in particular focusing on quasi-experimental designs. It defines the parameters of interest under the counterfactual approach, the assumptions and the identification strategies, presenting the Instrumental Variables (IV), Difference-in-Differences (DID) and the Synthetic Control (SC) methods.

Keywords: mediation, policy evaluation, direct effect, indirect effect, sequential conditional independence, quasi-experimental designs
JEL Classification: B41, C18, C21, C52, D04

*Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome. Email address: viviana.celli92@gmail.com

1 Introduction

In economics, causal analysis, or more in general program evaluation, is a fundamental instrument that allows to study causal effects of a variable of interest, known in literature as treatment. Causal analysis answers questions like: "Do subsidies to private capital boost firm's growth?" or "Are these effects positive or negative?". But this kind of analysis cannot answer to another important question: "Why are these treatments effective?". As pointed out by Gelman and Imbens (2013) not only the "effect of a cause", i.e. the treatment effect, seems relevant in many problems, but also "the cause of the effect", i.e. the mechanisms through which the total effect materializes. To use the words of Imai, Tingley and Yamamoto (2015): "A standard analysis of data [...] can only reveal that a program had such impacts on those who participated into it. It means that we can quantify the magnitude of these impacts, we can know how much a treatment affects an outcome, but these estimates tell us nothing about how. We know something about the causal effects, but nothing about causal mechanisms".

To overcome these limits a solution can be found in the causal mediation analysis, i.e. a formal statistical framework that can be used to study causal mechanisms. Following the definition given by Imai, Keele, Tingley, Yamamoto (2013) a mechanism is a process where a causal variable of interest, that is a treatment, influences an outcome through an intermediate variable, the mediator, that lies in the causal pathway between the treatment and the outcome variables. Studying causal mechanisms permits to know something more about social and economic policy implications than the total effect alone. This allows policy makers to optimize decisions, making them more efficient. The main fields in which mediation has been developed are psychology and sociology. For instance, Brader, Valentino and Suhay (2008) go beyond estimating the framing effects of ethnicity-based media cues on immigration preferences and ask: "Why the race of ethnicity of immigrants, above and beyond arguments about the consequences of immigration, drives opinion and behavior?". That is, instead of simply asking whether media cues influence opinion, they explore the mechanisms through which this effect operates. Consistent with earlier work suggesting the emotional power of group-based politics (Kinder and Sanders, 1996), the authors find that the influence of group-based media cues arises through changing individual levels of anxiety.

Another example is in electoral politics literature. Gelman and King (1990) found the existence of a positive incumbency advantage in the election. A few years later, in 1996, Cox and Kats lead the incumbency advantage literature in a new direction by considering possible causal mechanisms that explain why incumbents have an electoral advantage. They decomposed the incumbency advantage into a "scare off/quality effect" and effects due to other causal mechanisms such as name recognition and resource advantage.

Mediation is playing an increasing important role also in educational studies. Following the words of A. Gamoran "the next generation of policy research in education will advance if it offers more evidence on mechanisms so that the key elements of programs can be supported and the key problems in programs that fails to reach their goals can be repaired" (A. Gamoran, 2013,

President of the William T. Grant-Foundation). Also in a recent special issue of the Journal of Research on Educational Effectiveness focused on mediation and it has been noted that "such efforts in mediation analysis are fundamentally important to knowledge building, hence should be a central part of an evaluation study rather than an optional 'add-on'" (Hong, 2012). Can be found some empirical researches in the educational field like in Bijwaard and Jones (2018), who study the impact of education on mortality via cognitive ability, or Heckman, Pinto and Savelyev (2013), who study the effect of Perry Preschool Program through cognitive and non cognitive mechanisms.

Surprisingly, in economics mediation analysis has been much less contemplated, notwithstanding it has interesting and important implications. Few examples are in Simonsen and Skipper (2006), who evaluate the direct effect of motherhood's wage, Flores and Flores-Lagunes (2009), who evaluate the direct effect on earnings of the Job Corps program through work experience. Other contributions are given by Huber (2015), who used causal mediation framework to decompose the wage gap using data from the U.S. National Longitudinal Survey of Youth 1979, or by Huber, Lechner and Mellace (2017), who investigate whether the employment effect of more rigorous caseworkers in the counselling process of job seekers in Switzerland is mediated by placement into labor market programs. The common approach used to study causal mechanisms in economics is structural equation model (SEM), see for instance the seminal work by Baron & Kenny (1986). But, as demonstrated by Imai, Keele, Tingley and Yamamoto (2011), SEM is not the appropriate method to study and to identify causal mechanisms. They showed that structural models rely upon untestable assumptions and are often inappropriate even under the validity of those assumptions. In particular, conventional exogeneity assumptions alone are insufficient for identification of causal mechanisms[1], whereas it can be a sufficient condition for identification of the classical average treatment effect. In addition to that, the mediator could be interpreted as an intermediate outcome: in such a model we should control for a large set of covariates (pre and post treatment), risking to have different results depending on the covariates chosen and then increasing the sensitivity of the estimates. Therefore, the use of mediation in economics can be useful and efficient, and this is the main motivation of this brief exploration of mediation in economics.

To overcome these problems, relaxing the structural restrictions, over the last decades, some authors have moved mediation analysis in the potential outcome framework. Some examples are Robins and Greenland (1992); Pearl (2001); Petersen, Sinisi and van der Laan (2006); VanderWeele (2009); Imai, Keele and Yamamoto (2010); Hong (2010); Albert and Nelson (2011); Tchetgen Tchetgen and Shpister (2012); Vansteelandt, Bekaert and Lange (2012) from many others. As in the classical treatment analysis, using the counterfactual approach, rather than structural models, allows to formalize the concept of causality without making assumptions on the functional form of the parameters and, then, to have more flexible identification procedures.

[1] Structural models are misused also in the traditional causal analysis, because of the presence of strong assumptions to justify the causal interpretation of mathematical results. See for example James, Mulaik and Brett (1982), Pearl (1998) and many others.

Moreover, in this kind of models, it is not necessary to know the entire set of covariates that could affect the design. Most of this literature handles identification by assuming that the treatment and the mediator are conditionally exogenous given observed characteristics, an assumption known as Sequential Ignorability. Nevertheless, this assumption sometimes is hardly satisfied, above all in economics, because of the presence of post-treatment confounders, that can confound the relations between variables. To handle this problem, recently some researchers have used quasi-experimental designs inside the mediation framework. These procedures are particularly attractive in this context also because the gold standard of causal analysis, i.e. randomization of the treatment, is not a sufficient condition for the identification of causal mechanisms, a requirement that make the counterfactual approach more appropriate than structural models.
Causal mechanism is an important issue to better understand why a policy works and go beyond the limits of this approach is one of the aim of the current research fields. Mediation analysis seems to be one of the fittest frameworks to describe these relations and many researchers have developed new methods or have readapted the classical ones to go deep with the analysis. This is a promising methodology in economics because it permits to study causal mechanisms and to analyze the causal steps between treatment and outcomes and, then, it permits to give a causal interpretation to the changes that occur in between. In addition to that, these new methods that are emerging allow to do this kind of analysis without making too restrictive assumptions, a key issue in economic studies; mediation turns out to be a precious tool for policy makers. Thus, following the words of Imai, Keele and Yamamoto, also economics is trying to open his black box[2].
This paper reviews the methodological advancements in causal mediation literature in economics, in particular focusing on quasi-experimental designs, a recent perspective in the mediation panorama. The remainder of this paper is organized as follows: section 2 shows the counterfactual approach in mediation analysis and defines the parameters of interest; section 3 analyzes the assumptions required in mediation analysis; section 4 focuses on quasi-experimental designs, in particular showing instrumental variables (IV), difference-in-differences (DID) and synthetic control approaches; section 5 concludes.

2 Counterfactual approach

2.1 Definition of counterfactual mediation framework

Most recent research in mediation analysis uses counterfactual approach commonly exploited in causal inference, basing on the potential outcome framework proposed by Neyman (1923) for randomized experiments and then generalized to observational studies by Rubin (1974).

[2]Imai, K., L. Keele, D. Tingley, 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.

According to the main literature, formally I denote with D a binary treatment,[3] with M the mediator variable, that is assumed to have a boundary support and may be discrete or continuous, and with Y the outcome of interest. In this framework the potential outcome is defined as $Y(d', m)$ and the potential mediator is $M(d)$ with $d, d' \in \{0, 1\}$. I can write the realized outcome and mediator values as:

$$Y_i = D_i \cdot Y_i(1) + (1 - D)_i \cdot Y_i(0)$$
$$M_i = D_i \cdot M_i(1) + (1 - D)_i \cdot M_i(0)$$

where the subscripted i is the unit observation.
It is easy to see that for each unit i only one of the two potential outcomes or mediator states is observed. Thus, also in mediation analysis I have to face the so called missing values problem (Holland, 1986). Because of the presence of two driver variables we must also take into account the potential presence of an interaction between them, making the analysis more challenging.
The goal of mediation analysis is to decompose the total treatment effect of D on Y into the indirect and the direct effect. The first one reflects one possible explanation for why treatment works, explicitly defining a particular mechanism behind the causal impact and it answers the following counterfactual question: what change would occur to the outcome if the mediator changed from what would be realized under the treatment condition, that is $M_i(1)$, to what would be observed under the control condition, that is $M_i(0)$, while holding the treatment status at d?[4] The second one, the direct effect, represents all other possible explanations through which a treatment affects an outcome and it corresponds to the change in the potential outcome when exogenously varying the treatment but keeping the mediator fixed at its potential value $M_i(d)$. These methods attempt to assess what portion of the effect of the treatment operates through a particular intermediate variable and what portion operates through other mechanisms in order to prescribe better policy alternatives. Finally, mediation analysis is the set of techniques by which a researcher assesses the relative magnitude of these direct and indirect effects.

2.2 Definition of parameters

Using the potential outcome notation, I can define three quantities of interest, mostly used in mediation analysis, see for instance VanderWeele (2015):

1. CDE(m): $[Y_i(1, m) - Y_i(0, m)]$ is the controlled direct effect and it expresses how much the outcome would change between treated and control groups but keeping fix $M = m$. It quantifies the effect not mediated by M, but it is defined for every strata of m. If the effect

[3] Here I focus on binary treatment indicator for simplicity, but the methods can be extended easily to non-binary treatment, see for instance Imai, Keele & Tingley, 2010a.
[4] See for instance Keele, Tingley and Yamamoto, 2015

changes across vary level of m, then we are in presence of an interaction effect between D and M on Y.

2. NDE(d): $[Y_i(1, M(d)) - Y_i(0, M(d))]$ is the natural direct effect and it expresses how much the outcome would change if the treatment was exogenously set from 1 to 0 but, for each individual, the mediator was kept at the level it would have taken in treatment status equal d. It captures what the effect of the treatment on the outcome would remain if we were to disable the pathway from the treatment to the mediator.

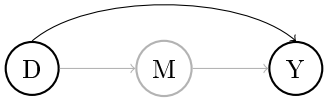

Figure 1: Natural direct effect

3. NIE(d): $[Y_i(d, M(1)) - Y_i(d, M(0))]$ is the natural indirect effect and it expresses how much the outcome would change if the treatment were set equal d but the mediator were changed from the level it would take if D=1 to the level it would take if D=0. It captures the effect of the treatment on the outcome that operates through the mediator.

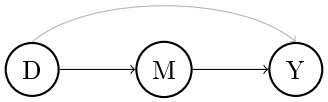

Figure 2: Natural indirect effect

These effects are defined at the unit level, implying that they are not observed for each observation i with the consequence that I cannot directly identify them without stronger assumptions. The reason is that they are defined with respect to multiple potential outcomes for the same individual and only one of those potential outcomes is observed in reality. So, I use the population averages for the identification of all the effects of interest. Basing on the potential outcome framework (Glynn 2012; Imai, Keele, Yamamoto 2010; Pearl 2001; Robins & Greenland 1992), I can identify these quantities of interest to disentangle the average total effect (ATE) given by $\Delta = E[Y(1) - Y(0)]$. First, I define the average indirect effect (ACME)[5] as:

$$\bar{\delta}(d) = E[Y(d, M(d)) - Y(d, M(1-d))] \qquad \forall\ d \in \{0, 1\} \tag{1}$$

[5] Also known as Average Causal Mediation Effect.

It corresponds to the change in mean potential outcome when exogenously shifting the mediator to its potential values under treatment and non treatment state but keeping the treatment fixed at $D = d$. Note that only one component of the right side equation is observable, whereas the other one is by definition unobservable (under treatment status d we never observe the value of M that it naturally would have under the opposite treatment state, i.e. $M(1-d)$).
In the same way, I define the average direct effect (ADE) as:

$$\bar{\theta}(d) = E[Y(d, M(d)) - Y(1-d, M(d))] \qquad \forall\ d \in \{0,1\} \tag{2}$$

It represents the average causal effect of the treatment on the outcome when the mediator is set to the potential value that would occur under treatment status d.
It can be easily shown that ATE can be rewritten as the sum of the natural direct and indirect effect defined on the opposite treatment status:

$$\begin{aligned}
\Delta &= E[Y_1 - Y_0] \\
&= E[Y(1, M(1)) - Y(0, M(0))] \\
&= E[Y(1, M(1)) - Y(0, M(1))] + E[Y(0, M(1)) - Y(0, M(0))] = \bar{\theta}(1) + \bar{\delta}(0) \\
&= E[Y(1, M(0)) - Y(0, M(0))] + E[Y(1, M(1)) - Y(1, M(0))] = \bar{\theta}(0) + \bar{\delta}(1)
\end{aligned}$$

I obtaine these results simply adding and subtracting the counterfactual quantity $E[Y(0, M(1))]$ after the second equality, and adding and subtracting $E[Y(1, M(0))]$ after the third equality. More in general, I can write this result as:

$$\Delta = \bar{\delta}(d) + \bar{\theta}(1-d) \qquad \forall\ d \in \{0,1\} \tag{3}$$

Obviously, neither effect is identified without further assumptions: only one of $Y(1, M(1))$ and $Y(0, M(0))$ is observed for any unit, because both outcomes cannot be observed at the same time as stated in the fundamental problem of causal inference; and the counterfactual quantities $Y(1, M(0))$ and $Y(0, M(1))$ are never observed for any individual, because I never observe the potential value of M defined under the opposite treatment state, but I only know the factual M that follows a particular treatment state. To face this identification issue I need to define a proper set of assumptions.

2.3 Controlled direct effect versus natural direct effect

An important advantage of the counterfactual notation is that it allows for the potential presence of heterogeneity. Such heterogeneity is important both in practical and theoretical, as it is often

the motivation for the endogeneity problems that concerns economists (Imbens and Wooldridge, 2009). In structural models the effects are assumed to be constant, implying that the effect of various policies could be captured by a single parameter. In mediation this heterogeneity is even more important, because it implies not only that the direct effect of the treatment on the outcome could be different across i, but also that this effect can be different for different values of the mediator. With the counterfactual notation, then, the presence of non linearities and interactions is not a problem, because I don't need to specify the functional form and I don't need to model the relations between variables. But if the effect of the treatment is the same for the entire population, meaning that it doesn't change for different level of the madiator, then there is no interaction between treatment and mediator. In this particular case, $CDE(m) = CDE(m')$, for $m \neq m'$, implying that the controlled direct effect is equal to the natural direct one, $CDE = NDE$ (Baron & Kenny, 1986). Formally:

$$\bar{\delta}(1) = \bar{\delta}(0) = \bar{\delta}$$
$$\bar{\theta}(1) = \bar{\theta}(0) = \bar{\theta}$$

In this situation, the difference between the total effect and the controlled direct effect gives the indirect effect, or more formally: $\Delta - \bar{\theta} = \bar{\delta}$.
Usually, in the empirical analysis the controlled direct effect and the natural direct effect do not coincide and then the difference between a total effect and a controlled direct effect does not generally give an indirect effect (Kaufman et al., 2004; Vanderweele, 2009) because there may simply be interaction between the effects of the exposure and mediator on the outcome, not guaranteeing the additional linearity functional form of the effects.

3 Assumptions

3.1 Classical Assumptions

Usually, in economics I can't manage a controlled experiment. In this situation I must rule out the presence of confounders. But, in mediation analysis, because of the particular structure of the variables' relations, it is important to point out what kind of confounders I have to face.
Consider a classical mediation framework, in which X is a set of pre-treatment observable covariates and W is a set of post-treatment observable confounders, like in figure 3.

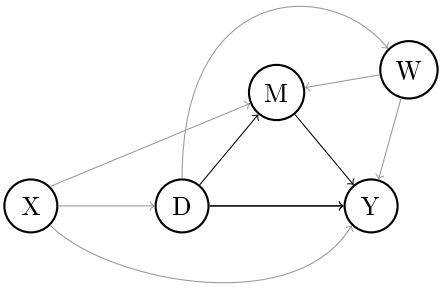

Figure 3: Mediation framework with: D=treatment; M=mediator; Y=outcome; X=pre-treatment confounders; W=post-treatment confounders

For example, suppose I want to assess whether a program of subsidies (D) increases firms' productivity (Y) and whether the share invested in R&D (M) may mediate part of this effect. In this example, investments in R&D may be a potential mediator because is affected by subsidies treatment and in turn may affect, at least partially, productivity outcome. But to interpret this association as a causal effect, I need to think carefully about and try to control for variables that may be confounders of the treatment-outcome relationship (X) and/or of the mediator-outcome relationship (W). For example, there might be a firms' size or firms' performance variables (X) that affect the participation in the program (D) and the firms' productivity (Y) or other factors, such as quality of administration or the presence of a network (W), that affect both the level of investments in R&D and productivity. It is important to note that these W confounders could be affected by the treatment itself.
In such a context, I need to distinguish two situations: the identification of controlled effects and identification of natural effects. Following VanderWeele (2015) to estimate the CDE we need two assumptions:

- A1. There must not be confounders between treatment and outcome relationship
- A2. There must not be confounders between mediator and outcome relationship

For the satisfaction of the first assumption is sufficient randomize the treatment, but even with randomized treatment the second assumption might not hold. If I refer to the previous example, to satisfy A1 I need to adjust for common causes of the exposure and the outcome - for example information about firms' size or firms' performance or any other factor (X) that can confound this relation in the analysis; or I can give subsidies randomly, implying the same distribution of X for treated and non-treated firms. At the same time, to satisfy A2 I need to adjust for common causes of the mediator-outcome relation - for example information about administration's quality or other factors (W) that can confound this relation. In this case, I need to think carefully to all possible post-treatment confounders and include them in the analysis, because the randomization

of the treatment is not a sufficient condition to control for W.
To identify natural direct and indirect effects I need two more assumptions. In particular:

- A3. There must not be confounders between treatment and mediator relationship
- A4. There must not be confounders affected by the treatment between mediator and outcome relationship

Also in this case, to satisfy A3 is sufficient randomize the treatment, but again for the fourth assumption this is not enough. In particular, A4 is a strong assumption, because it requires that there is nothing on the pathway from the treatment to the mediator that also affects the outcome. This assumption is more plausible if the mediator occurs shortly after the treatment (VanderWeele and Vansteelandt, 2009). If I consider again the previous example, the size of the firm could be a confouder of the treatment-mediator relation and then it must be included in the set of covariates (X) or I can randomly assign D. But if I consider A4, I have to take into account possible factors that could be affected by the treatment and that in turn affect the mediator and/or the outcome. For example, firms that receive subsidies could have more benefits (tax, bureaucratic) that could in turn affect R&D investments and productivity. The problem with this assumption is that, even if I have in mind these factors before the analysis, I can't have the exact measure of these confounders, because I don't know before the value that they will assume after the treatment.
Another example could be the effect of a job training program (D) on the probability to find a job (Y). It could be possible that the program is designed in two steps: the first part of the program in which I can find different activities and a second part in which there is, for example, a PC course (M). In this kind of design, I can study how much the probability to find a job increases thanks to the PC course and/or thanks to the other components of the job training program. Also in this case, to correctly identify the direct and the indirect effect I must be sure to satisfy the previous four assumptions. In other words, I have to control for all possible pre-treatment confounders, like gender, age, education, kind of job, how long the individual is unemployed and so on, and I have to control for all possible post-treatment confounders, eventually affected also by the treatment, like the previous knowledge of PC, attitude and so on.
It is important to note that assumptions 1-4 implicitly imply an assumption of temporal ordering (Cole & Maxwell, 2003). If the temporal ordering assumptions were not satisfied, then neither would the no unmeasured confounding assumptions, and then the association would not represent causal effect. For this last assumption it is important to use panel data to measure the various factors at different time: such framework consisting of an initial treatment, an intermediate mediator, a final outcome and, possibly, observed covariates. Differently, with cross sectional data, I cannot determine the direction of causality or the relative magnitude of the two possible directions that causality may operate and I cannot distinguish between mediation and confounding (see for instance, Baron & Kenny, 1986). In addition, it is important to have in mind a strong theory to give the right causal interpretation,. Another issue is that when the treatment, the

mediator and the outcome vary over time, we have to control for the prior values of these variables to make no confounding assumptions more plausible and to rule out the possibility of reverse causality (T. J. Vanderweele, 2015): even if I know the temporal ordering, it is possible that prior values of the variables serve as the most important confounding variables.

3.2 Identification under Sequential Ignorability

The key insight is that under randomized designs ATE is identified, but direct and indirect effect are not. Even in the presence of a double randomization of the treatment and the mediator the effects of interest are not identified without further assumptions. In fact, even if both treatment and mediator are exogenous, and then the conventional exogeneity assumption is satisfied, simply combining the effect of T on M and the effect of M on Y is not sufficient for the identification of the indirect effect. The assumption called "Sequential Ignorability" is a partial solution to this problem and so far the most used. There are different interpretations of this assumption, with different implications and different formalizations. The most used version and maybe the most flexible is the one given by Imai, Keele and Yamamoto (2010). Formally, it is expressed as:

$$\{Y_i(d', m), M_i(d)\} \perp D_i | X_i = x \tag{4}$$

$$Y_i(d', m) \perp M_i(d) | D_i = d, X_i = x \tag{5}$$

where:

$$Pr(D_i = d | M_i = m, X_i = x) > 0 \tag{6}$$

$$\forall\ d \in \{0, 1\} \text{ and } m, x \text{ in the support of } M, X^6$$

The first part of the sequential ignorability assumption, equation (4), is the classical conditional independence of the treatment, also known as no-omitted variable bias, conditional exogeneity or unconfoundedness, see for instance Imbens (2004). By equation (4), there are no unobserved confounders jointly affecting the treatment and the mediator and/or the outcome given X, meaning that I can consistently identify the effect of D on Y and D on M. In non-experimental designs, the validity of this assumption hinges on the richness of pre-treatment covariates, whereas in experimental designs, this assumption holds if the treatment is either randomized within strata defined by X or randomized unconditionally[7]. The second part of sequential ignorability assumption, equation (5), states that there are no unobserved confounders jointly affecting the mediator and the outcome once I condition on D and X. It means that there are no unobserved confounders between mediator and outcome, ruling out the presence of post-treatment confounders not captured by X. This is a strong assumption because randomizing both treatment and

[6] Imai, Keele, Tingley and Yamamoto (2011) wrote this common support assumption as: $0 < Pr(D_i = d | X_i = x)$ and $0 < P(M_i = m | D_i = d, X_i = x)$ for $d = 0, 1$ and all x and m in the support of X and M.

[7] In this case, the stronger version of the assumption $\{Y_i(d', m), M_i(d), X\} \perp D_i$ is satisfied.

mediator does not suffice for this assumption to hold; in addition to this, it is more plausible if treatment and mediator are measured at a short distance, as I mentioned in the previous subsection. The last part of sequential ignorability, equation (6), is the common support assumption. It states that the conditional probability to receive or not receive the treatment given M and X, recalling the propensity score literature, is larger than zero[8]. By Bayes' theorem, this version of common support implies that $Pr(M_i = m|D_i = d, X_i = x) > 0$ if M is discrete or that the conditional density of M given D and X is larger than 0 if M is continuous. The main implication of the equation (6) is that conditional on X, the mediator state must not be a deterministic function of the treatment, otherwise no comparable units in terms of the mediator are available across different treatment states (Huber, 2019). In other words, there must be different values of M once I condition on D and X, in order to compare different mediator states inside the same group defined by the treatment status. Under sequential ignorability (equations 4-6), it is possible to identify causal mechanisms, in particular, I can get the nonparametric identification of the counterfactual quantity $E[Y_i(d, M(d')|X_i = x]$, proved by Imai, Keele and Yamamoto (2010), implying the nonparametric identification of the average natural direct (ADE) and the average natural indirect effect (ACME). In the standard causal mediation analysis the nonparametric identification of the counterfactual quantity is the following:

$$
\begin{aligned}
&E[Y_i(d, M_i(d'))|X_i = x] = \\
&= \int E(Y_i(d, m)|M_i(d') = m, X_i = x]\, dF_{M_i(d')|X_i=x}(m) \\
&= \int E(Y_i(d, m)|M_i(d') = m, D_i = d', X_i = x]\, dF_{M_i(d')|X_i=x}(m) \\
&= \int E(Y_i(d, m)|D_i = d', X_i = x]\, dF_{M_i(d')|X_i=x}(m) \\
&= \int E(Y_i(d, m)|D_i = d, X_i = x]\, dF_{M_i(d')|D_i=d',X_i=x}(m) \\
&= \int E(Y_i(d, m)|M_i = m, D_i = d, X_i = x]\, dF_{M_i(d')|D_i=d',X_i=x}(m) \\
&= \int E(Y_i|M_i = m, D_i = d, X_i = x]\, dF_{M_i(d')|D_i=d',X_i=x}(m) \\
&= \int E(Y_i|M_i = m, D_i = d, X_i = x]\, dF_{M_i|D_i=d',X_i=x}(m)
\end{aligned}
$$

where, assuming a continuous mediator, the first equality follows from the law of iterated expectation; equation (4) is used to establish the second, the fourth and the last equalities; equation (5) is used to establish the third and the fifth equalities, whereas the sixth equality follows from the fact that $M_i = M_i(D_i)$ and $Y_i = Y_i(D_i, M_i(D_i))$, also known as observational rule (T. VanderWeele, 2015) or consistency assumption (Imai, Keele, Tingley and Yamamoto, 2011).

[8] In the classical causal analysis to identify the ATE I face the weaker common support assumption: $Pr(D_i = d|X_i = x) > 0$

Sequential ignorability used in the counterfactual analysis is crucially different w.r.t. the classical exogeneity assumption used in the structural models. In particular, as I said before, to identify causal mechanisms, and then the indirect effect that goes from T to Y through M, is not sufficient randomize both treatment and mediator. Differently, if I use structural models, it is required to satisfy only the exogeneity assumption, meaning that it's sufficient the double randomization of T and M. Nevertheless, the resulting estimation is consistent only if there is not heterogeneity effect. In particular, in the first case I can identify the causal mediation effect (T $\rightarrow$ M $\rightarrow$ Y) in which I am interested in, whereas in the second case I can just identify the causal effect of the mediator (T $\rightarrow$ M and M $\rightarrow$ Y). These two quantities coincide only in the absence of heterogeneity. Under exogeneity assumption and in the absence of heterogeneity, then, I can consistently estimate only CDE, because in this particular case this quantity is equal to NDE. The interesting fact is that, in the presence of heterogeneity, the exogeneity assumption still holds if treatment and mediator are randomized, but the correlation between the error terms of M and Y is different from 0, implying biased estimations of the effects, that structural models are not able to capture.

3.3 Other interpretations of Sequential Ignorability

The main limit of this result is that the nonparametric identification works only if I don't condition on post-treatment confounders, implying that the set of pre-treatment observable confounders must be sufficient to control for them, requirement not always credible. This issue has been addressed by Robins (2003). In his fully randomized causally interpreted structured tree graph model (FRCISTG), he used a different version of sequential ignorability: the first part is the same of equation (4), whereas equation (5) is replaced by $Y_i(d', m) \perp M_i(d) | D_i = d, Z_i = z, X_i = x$, where Z is a vector of post-treatment confounders. This is an important practical advantage because permits to control for observable variables that could confound the relationship between the mediator and the outcome. But it comes at the cost of adding the parametric assumption of non-interaction between direct and indirect effect: $Y_i(1, m) - Y_i(0, m) = B_i$, where B_i is a random variable independent of m. This condition has two implications: (i) absence of heterogeneity; (ii) the same value of the direct effect regardless the level of the mediator, i.e. the independence between the direct and the indirect effect. Therefore, it exists an important trade-off: if I condition on post-confounders, I need to assume a non-interaction assumption to identify natural effects, which is very restrictive condition and it doesn't permit a nonparametric identification. On the other hand, if I don't condition on post-treatment confounders, but assuming that all the X's are sufficient to control for them, I can identify the effects nonparametrically without any parametric restrictions.

Another formalization of Sequential Ignorability is given by Pearl (2001). In particular, in his Theorem 1 and Theorem 2 for the identification of the average natural direct effect and in Theorem 4 for the identification of the average natural indirect effect, he used a different set of assumptions arriving anyway at the same expression of ADE and ACME given by Imai,

Keele and Yamamoto (2010). It is important to note that sequential ignorability implies Pearl's assumptions, whereas the converse in not always true, but in practice, the difference is only technical. Another advantage of sequential ignorability is that it is easier to interpret than Pearl's assumptions, in which I have an independence between two potential quantities[9]. This difficulty in the interpretation is pointed out also by Pearl himself: "Assumptions of counterfactual independencies can be meaningfully substantiated only when cast in structural form"[10]. In contrast, in the second part of sequential ignorability, eq. (5), I have the observed value $M_i(d)$ independent of potential outcome, in other words M_i is effectively randomly assigned given $D_i = d$ and $X_i = x$, a concept that is easier to understand.
A further version of sequential ignorability is given by Petersen, Sinisi and Van der Laan (2006). They split equation (4) into two parts: $Y_i(d, m) \perp D_i | X_i = x$ and $D_i \perp M_i(d) | X_i = x$, whereas equation (5) is the same[11]. This is just a mathematical difference, because in experimental designs, in which treatment is randomized, equation (4) is equivalent to them. To identify the natural direct effect they also assume that the potential value of mediator under non-treatment state is independent of the potential outcome. Formally, $E[Y_i(d, m) - Y_i(0, m) | M_i(0) = m, X_i = x] = E[Y_i(d, m) - Y_i(0, m) | X_i = x]$, meaning that the potential value of the mediator under non treatment state, $M_i(0)$, doesn't give any additional information on the effect of the treatment. This additional assumption is necessary to identify the counterfactual quantity $Y(d, M(0))$. Anyway, if treatment is randomized this last assumption is not necessary for the nonparametric identification given by Imai, Keele and Yamamoto (2010), making their sequential ignorability a preferable solution once again.

4 Quasi-experimental designs

As mentioned in the previous section, most recent research in mediation analysis considers more general identification approaches based on the potential outcome framework, commonly used in treatment evaluation (Rubin, 1974) to overcome the limits of structural models. The gold standard of this approach is the randomness of the treatment, a condition that is easily met in experiments. When treatment or mediator cannot be determined exogenously, the only way to estimate the parameters of interest and give them a causal interpretation is to use quasi-experimental designs, in which endogeneity can be controlled under particular assumptions. Mediation analysis borrowed these methods from causal literature in order to identify and estimate causal mechanisms, but, nowadays, there are only few studies using these approaches. Can be found some examples in Instrumental variables (See for example Robins and Greenland, (1992); Geneletti (2007); Imai et el. (2013); Powdthavee et al. (2013); Burgess et al. (2015); Jhun (2015); Frölich and Huber (2017)), Difference-in-differences (see Deuchert, Huber and Schelker

[9] The assumption given in Pearl(2001) is: $Y_i(d', m) \perp M_i(d) | X_i = x$

[10] See Pearl (2001), pag. 416

[11] In particular, they use $Y_i(d, m) \perp D_i | X_i = x$ and $D_i \perp M_i(d) | X_i = x$ to identify controlled direct effect and they add equation (5) to identify natural direct effect.

(2018); Huber and Steinmayr (2017)) and synthetic control (see Mellace and Pasquini (2018)), while, at the best of my knowledge, there are not still studies using regression discontinuity design[12]. In the next section I will discuss some of them.

4.1 Instrumental variables

Recently, part of the literature tried to study causal mechanisms through instrumental variables (IV) methods (see Robins and Greenland (1992); Imai et al. (2013) from many others). The reason is that, in some empirical applications, sequential ignorability is not a credible assumption to rule out the presence of post-treatment confounders and an instrument could be an important tool to solve the problem of the mediator's endogeneity. In other cases, also the treatment is not exogenous even after conditioning on a set of pre-treatment covariates and a second instrument could be used for this kind of endogeneity. Can be found two different ways in which mediation analysis with IV has been dealt. Some authors identified direct and indirect effects through structural models. For example, Powdthavee, Lekfuangfu and Wooden (2013) studied the impact of education on subjective well-being (SWB) through the mediator income. They used different timing of education laws across states of Australia and shocks in personal income (such as lottery wins etc.) as instruments respectively of treatment and mediator. Assuming the independence between instruments, they estimate the direct and indirect effect using the structural equation model (see Baron & Kenny, 1986), inside a 2SLS framework. Other studies used two instruments and a parametric identification such as Burgess et al. (2015) and Jhun (2015). Ten Have et al. (2002) used treatment-covariates interactions as instruments for the mediator, but imposing the absence of the treatment-mediator, mediator-covariate and treatment-covariate interactions in the outcome model, implying an identification based on strong structural restrictions. The limit of this structural methods is that they don't allow for the existence of a heteorgeneous effect between direct and indirect effect.

The second way in which mediation analysis with IV can be studied is using the potential outcome framework. An important contribution is given by Chen, Chen and Liu (2017), who studied the gender of the second born on the first born education outcome, through the sibling size (also interpreted as fertility choice) mediator. In their study, they assume a randomized sibling gender and they use a twinning indicator at the second birth as instrument for the mediator (following the studies of Rosenzweig and Wolpin (1980); Black, Devereux and Salvanes (2005); Angrist, Lavy and Schlosser (2010)). Their IV estimates give a causal interpretation limited only to complying families, whose sibling size would rise with twinning at the second birth, i.e. $M(Z = 1) > M(Z = 0)$, but, on the other hand, allowing for heterogeneous effect, i.e. interaction between treatment and mediator. In particular, they found that having a younger brother lowers the potential sibling size of a first-born girl to a degree that the positive indirect effect cancels out the negative direct effect on her education outcomes, resulting in a near zero

[12] See M. Angelucci, V. Di Maro (2010): they provide a practical guide for the identification of treatment effect on eligibles and the indirect effect on ineligibles based on conditional independence, RD and IV assumptions

total effect. These results offer new evidence about gender bias in family settings that has not been detected in the previous literature. This was possible thanks to the decomposition of the total effect and thanks to the presence of heterogeneity captured by the interaction between sibling size and sibling gender. A second contribution using the potential outcome approach is given by Frölich and Huber (2017). They used a counterfactual framework and join a nonparametric identification using two different instruments respectively for treatment and mediator, allowing, then, for the endogeneity of them. In addition, both instruments and mediator can be discrete or continuous. The main advantage of their result is that they identify natural and controlled effects for all treatment compliers, overcoming the limit of identification only of the controlled direct effect for subpopulations defined on compliance in either endogenous variable (see Miquel, 2002). They applied this method on two empirical studies. One of them is about the effect of education on the social life outcome through income. Treatment is instrumented by an increase in the UK minimum school leaving age in 1971 from 15 to 16 years (see also Oreopoulos, (2006) and Brunello et al., (2013)), whereas the annual individual income is instrumented by windfall income (Lindhal, (2005) and Gardner and Oswald, (2007)). They found a positive effect of education on social life functioning, but disentangling the total effect on compliers (LATE) showed a positive direct effect, whereas the indirect effect is close to 0 and not significant. They then conclude that education affects social functioning, but through different mechanisms than income (Huber and Frölich, 2017).

4.2 Difference-in-differences

The first contribution that deals the identification of direct and indirect effect using a different framework than sequential ignorability and instrumental variables approach is given by E. Deuchert, M. Huber and M. Schelker (2018). They disentangle the total effect basing on a difference-in-differences (DID) approach within subpopulation or strata (Frangakis and Rubin, 2002) defined upon the reaction of a binary mediator to treatment, implying the presence of four subpopulations: always takers, never takers, compliers and defiers (see for instance Angrist, Imbens and Rubin, 1996). In particular, they identify the direct effect on always takers and never takers, whose mediator doesn't react to treatment, i.e. treatment doesn't change the mediator's state, corresponding to the controlled direct effect, and then they identify the indirect effect and the direct effect on compliers, whose mediator reacts to treatment. The main assumptions that they use are the classical random treatment assignment; the second one is the monotonicity assumption that comes from the local average treatment effect (LATE) literature (see Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996), ruling out the presence of defiers. The last important set of assumptions is the common trend assumptions, which come from the DID literature, but now defined across strata. This fact permits to control for post-treatment confounders and it allows for differences in the effects of unobservable confounders on specific potential outcomes across strata, as long as these differences are time constant. As discussed in this paper, the identification of the effects of interest under principal strata in mediation has been criticized for not permitting

a decomposition of direct and indirect effect on compliers in a DID framework and focussing on subgroups that may be less interesting than the entire population (VanderWeele 2008). But thanks to previous set of assumptions the authors identify the effects on compliers and they present an empirical application in which the effect on subgroups is relevant for political decision making[13]. A second critique is about confusion made in the literature between mediation and principal stratification causal effects (VanderWeele 2012). In particular, it is important to note that $E[Y_1 - Y_0|M(1) = 1, M(0) = 0]$ is the total causal effect of treatment on the outcome for the compliers subgroup and it doesn't always correspond to the mediated effect. To notice this fact, can be observed that this effect can be nonzero even if the intermediate variable has no effect on the outcome, meaning that M is not a mediator. This happens whenever M is a surrogate for the effect of the treatment on the outcome: surrogacy concerns whether the effect of a treatment on an outcome can be predicted by the effect of a treatment on an intermediate variable, whereas mediation concerns whether the effect of treatment goes to the outcome through the mediator. A good surrogate may be often a mediator, but it need not be (Vander Weele, 2012). Principal stratification is a good framework to capture surrogacy, whereas natural effects (Pearl 2001, from many others) are the appropriate concept to study mediation. An intuitive example is given by Lindsay Page (2012), who provides evidence that Career Academies program (D) had a substantial effect on subsequent earnings (Y) those for whome the program would change exposure to the world-of-work (M) but not those for whome it would not change exposure to the world-of-work. In her analysis, she used a Bayesian approach to principal stratification and she used covariates to attempt to predict which principal stratum different individuals belong to. But, even if these assumptions hold, it could happen that there are still some unmeasured confounders of the mediator-outcome relationship, like motivation (U), that make M a surrogate rather than a mediator, like in Figure 4.

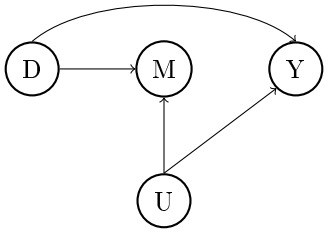

Figure 4: Principal strata causal effect with no mediation

A possible solution is to study mediation with principal strata approach, but adding the sequential ignorability assumption to rule out the potential presence of post-treatment confounders.

[13]The empirical application is about the Vietnam draft lottery in the US (1969-1972) on political preferences and personal attitudes. The mediator of interest is military service during the Vietnam War.

4.3 Synthetic control

To the best of my knowledge, the only contribution that uses synthetic control method (SCM) to study causal mechanisms is given by G. Mellace and A. Pasquini (2018). The main advantage of this method is that it estimates total causal effects, even in presence of only one treated unit and few control units (Abadie and Gardeazabal (2003)). They develop a generalization of SCM that allows disentangling the total effect into its direct and indirect component defining a Mediation Analysis Synthetic Control (MASC). The procedure that they use consists in re-weighting control unit post-intervention outcomes by choosing weights that minimize the distance between treated and control in pre-intervention observable characteristics as well as in post-intervention values of mediator. This allows to mimic what would have happened to the treated in the absence of the intervention if her mediator were set to her potential mediator under treatment (Mellace and Pasquini, 2018). In particular, they use a dynamic factor model with interactive fixed effects as in Abadie et al. (2010).

5 Conclusion

Mediation analysis is a promising methodology in economics, because it allows to study causal mechanisms of transmission of a policy without making unreliable and often restrictive assumptions: it permits to know, not only if a policy is working or not, but also why, going into a deeper level of analysis. In literature, there are not many economic applications and it could be due to technical difficulties and to the absence of clear methodological developments. I reviewed the pillars of this methodology, presenting current results and advancements and providing that it's a validated method, that can be used to investigate the changes that occur between inputs and outputs, answering the opened questions of economic studies. Causal mediation analysis is the statistical tool to understand causal mechanisms and it may bring to an improvement in the power of quantitative analysis of economic phenomena.

This paper provides a survey of methodological developments in causal mediation analysis in economics, with a specific focus on quasi-experimental designs. I presented several methods, often used by economists and statisticians, that are clearly useful and fruitful for economic causal analysis. In the first part, I defined direct and indirect effects, both formally and mathematically. Next, I discussed the main assumptions needed for the identification of the counterfactual quantities of interest, with particular attention to the sequential ignorability assumption. In the fourth section I reviewed the main studies that use quasi-experimental designs, a new frontier in this field, discussing in particular instrumental variables, difference-in-differences and synthetic control approaches.

References

[1] Abadie, A., A. Diamond, J. Hainmueller (2010): "Synthetic control methods for comparative case studies: estimating the effect of california's tobacco control program", *Journal of the american statistical association*, 105(490), 493-505.

[2] Abadie, A., J. Gardeazabal (2003): "The economic costs of conflict: a case study of the Basque Country", *American economic review*, 93(1), 113-132.

[3] Albert, J. M., S. Nelson (2011): "Generalized causal mediation analysis", *Biometrics*, 67, 1028-1038.

[4] Angelucci, M., V. Di Maro (2010): "Program evaluation and spillover effects", *Working paper, University of Michigan.*

[5] Angrist, J., G. Imbens, D. Rubin (1996): "Identification of causal effects using Instrumental Variables", *Journal of American statistical association*, 91, 444-472.

[6] Angrist, J., Lavy V., Schlosser A. (2010): "Multiple experiments for the causal link between the quantity and quality of children", *Journal of labor economics*, 28(4), 773-824.

[7] Baron R.M., Kenny D.A. (1986): "The moderator-mediator variable distiction in social psychological research: conceptual, strategic and statistical considerations", *Journal of personality and social psychology*, 51, 1173-1182.

[8] Bijwaard, G. E., A. M. Jones (2018): "An IPW estimator for mediation effects in hazard models: with an application to schooling, cognitive ability and mortality", *Empirical economics*, pp. 1-47.

[9] Black, S. E., P. J. Devereux, K. J. Salvanes (2005): "The more the merrier? The effect of family size and birth order on children's education", *The quarterly journal of economics*, 120(2), 669-700.

[10] Brader, T., N. A. Valentino and E. Suhay (2008): "What triggers public opposition to immigration? Anxiey, group cues and immigration", *American Journal of Political Sciences*, 52(4), 959-978.

[11] Burgess, S., R. M. Daniel, A. S. Butterworth, S. G. Thompson (2015): "Network mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways", *International journal of epidemiology*, 44, 484-495.

[12] Cerqua, A., G. Pellegrini (2013): "Do subsidies to private capital boost firm's growth? A multiple regression discontinuity design approach", *Journal of public economics.*

[13] Chen, S. H.,Y. C. Chen, J. T. Liu (2017): "The impact of family composition on educational achievement", *forthcoming in the journal of human resources.*

[14] Cole, D. A., S. E. Maxwell (2003): "Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling", *Journal of abnormal pshychology*, 112, 558-577.

[15] Cox, G. W., J. N. Kats (1996): "Why did the incumbency advantage in U.S. House elections grow?", *American journal of political science*, 20(2), 478-497.

[16] Deuchert, E., M. Huber, M. Schelker (2017): "Direct and indirect effects based on difference-in-differences with an application to political preferences following the Vietnam draft lottery", *Journal of Business & Economic statistics*, 1537-2707.

[17] Flore C. A., A. Flores-Lagunes (2009): "Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness", *IZA DP No. 4237*.

[18] Frangakis, C., D. Rubin (2002): "Principal stratification in causal inference", *Biometrics*, 58, 21-29.

[19] Frölich, M., M. Huber (2017): "Direct and indirect treatment effect - causal chains and mediation analysis with instrumental variables", *Journal of the royal statistical society: series B*, 79(5), 1645-1666.

[20] Gardner, J., A. J. Oswald (2007): "Money and mental wellbeing: a longitudinal study of medium-sized lottery wins", *Journal of health economics*, 26(1), 49-60.

[21] Gelman, A., G.W. Imbens (2013): "Why ask Why? Forward causal inference and reverse causal questions", *NBER Working paper No. 19614*.

[22] Gelman, A., G. King (1990): "Estimating incumbency advantage without bias", *American Journal of Political Science*, 34(4), 1142-64.

[23] Glynn, A. N. (2012): "The product and difference fallacies for indirect effects", *Journal of political science*, 56, 257-269.

[24] Havelmo, T. (1943): "The statistical implications of a system of simultaneous equations", *Econometrica*, 11, 1-12.

[25] Heckman, J., R. Pinto, P. Savelyev (2013): "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes", *American economic review*, 103, 2052-2086.

[26] Holland, P. W. (1986): "Statistics and causal inference", *Journal of the american statistical association*, 81, 945-60.

[27] Hong M. (2010): "Ratio of mediator probability weighting for estimating natural direct and indirect effects", in *Proceedings of the American statistical association, Biometrics section*, p. 2401-2415. Alexandria, VA: American Statistical Association.

[28] Hong M. (2012): "Editorial comments", *Journal of educational effectiveness*, 5, 213-214.

[29] Huber, M. (2015): "Causal pitfalls in the decomposition of wage gap", *Journal of business and economic statistics*, 33, 179-191.

[30] Huber, M. (2019): "A review of causal mediation analysis for assessing direct and indirect treatment effects", *Working paper No. 500.*

[31] Huber, M., M. Lechner, 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, 180-183.

[32] Imai, K., L. Keele, D. Tingley, 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.

[33] Imai, K., L. Keele, T. Yamamoto (2010): "Identification, inference and sensitivity analysis for causal mediation effects", *statistical science*, 25, 51-71.

[34] Imai, K., D. Tingley, T. Yamamoto (2013): "Experimental designs for identifying causal mechanisms", *Journal of the Royal statistical society, Series A*, 176, 5-51.

[35] Imbens, G. W. (2004): "Nonparametric estimation of average treatment effect under exogeneity: a review", *The review of economics and statistics*, 86, 4-29.

[36] Imbens, G. W., J. Angrist (1994): "Identification and estimation of local average treatment effects", *Econometrica*, 62, 467-475.

[37] Imbens, G. W., J. M. Wooldridge (2009): "Recent developments in the econometrics of program evaluation", *Journal of economic literature*, 47, 5-86.

[38] Lindsay, C. Page (2012): "Principal stratification as a framework for investigating mediational processes in experimental sttings", *Journal of Research on educational effectiveness*, 5(3), 215-244.

[39] Kaufman, J. S., R. F. Maclehose, S. Kaufman (2004): "A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation", *Epidemiologic Perspectives & innovations*, 1, 4.

[40] Keele, L., D. Tingley, T. Yamamoto (2015): "Identifying mechanisms behind policy interventions via causal mediation analysis", *Journal of policy analysis and management*, 34, 937-963.

[41] Kinder, D.R., L. Sanders (1996): "Divided by color: racial politics and democratic ideals", *Chicago: University of Chicago Press.*

[42] Kopmans, T. C., H. Rubin, R. B. Leipnik (1950): "Measuring the equation systems of dynamic economics". In T. C. Koopmans (Ed.), statistical inference in dynamic economic models, Cowles Commission monograph 10. New York: Wiley, 1950. 53-237.

[43] Mackinnon, D. (2008): "Introduction to statistical mediation analysis", New York: Routledge.

[44] Mellace, G., A. Pasquini (2018): "Mediation analysis synthetic control", *mimeo*.

[45] Miquel, R. (2002): "Identification of dynamic treatment effects by instrumental variables", *University of St. Gallen economics discussion paper series*, 2002-11.

[46] Oreopoulos, P. (2006): "Estimating average and local treatment effects of education when compulsory schooling laws really matter", *American economic review*, 96(1), 152-175.

[47] Pearl, J. (2001): "Direct and indirect effcts", in *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pp. 411-420, San Francisco. Morgan Kaufman.

[48] Petersen, M. L., S. E. Sinisi, M. J. Van der Laan (2006): "Estimation of direct causal effects", *Epidemiology*, 17, 276-284.

[49] Powdthavee, N., W. N. Lekfuangfu, M. Wooden (2013): "The marginal income effect of education on happiness: estimating the direct and indirect effect of compulsory schooling on well-being in Australia", *IZA discussion paper No. 7365*.

[50] Robins, J. M. (2003): "Semantics of causal DAG models and the identification of direct and indirect effects", in *In highly structured stochastic systems*, ed. by P. Green, N. Hjort and S. Richardson, 70-81, Oxford. Oxford University Press.

[51] Robins J. M., S. Greenland (1992): "Identifiability and exchangeability for direct and indirect effects", *Epidemiology*, 3, 143-155.

[52] Rosenzweig, M. R., K. Wolpin (1980): "Testing the quantity-quality fertility model: the use of twins as a natural experiment", *Econometrica*, 48(1), 227-240.

[53] Rubin, B. (1974): "Estimating causal effects of treatment in randomized and non nonrandomized studies", *Journal of educational psychology*, 66, 688-701.

[54] Shadish, W. R., T. D. Cook, D. T. Campbell (2001): "Experimental and quasi-causal designs for generalized causal inference", Boston, Houghton Mifflin.

[55] Simonsen, M., L. Skipper (2006): "The costs of motherhood: an analysis using matching estimators", *Journal of applied econometrics*, 21, 919-934.

[56] Tchetgen Tchetgen, E. J., I. Shpitser (2012): "Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis", *The annals of statistics*, 40, 1816-1845.

[57] VanderWeele, T. J. (2008): "Simple relations between principal stratification and direct and indirect effects", *Statistics & Probability letters*, 78, 2957-2962.

[58] VanderWeele, T. J. (2009): "Marginal structural models for the estimation of direct and indirect effects", *Epidemiology*, 20, 18-26.

[59] VanderWeele, T. J. (2012a): "Comments: should principal stratification be used to study mediational processes?", *Journal of research on educational effectiveness*, 5(3), 245-249.

[60] VanderWeele T.J. (2015): "Explanation in causal inference. Methods for mediation and interaction", Oxford University Press.

[61] VanderWeele, T. J., S. M. Vansteelandt (2009): "Conceptual issues concerning mediation, interventions and composition", *Statistics and its inference*, 2, 457-468.

[62] Vansteelandt, S., M. Bekaert, T. Lange (2012): "Imputation strategies for the estimation of natural direct and indirect effects", *Epidemiologic Methods*, 1, 129-158.