CESifo WORKING PAPERS

7855
2019

Original Version: September 2019
This Version: October 2020

Direct and Indirect Effects Based on Changes-in-Changes

Martin Huber, Mark Schelker, Anthony Strittmatter

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7855*
Category 12: Empirical and Theoretical Methods

Direct and Indirect Effects Based on Changes-in-Changes

Abstract

We propose a novel approach for causal mediation analysis based on changes-in-changes assumptions restricting unobserved heterogeneity over time. This allows disentangling the causal effect of a binary treatment on a continuous outcome into an indirect effect operating through a binary intermediate variable (called mediator) and a direct effect running via other causal mechanisms. We identify average and quantile direct and indirect effects for various subgroups under the condition that the outcome is monotonic in the unobserved heterogeneity and that the distribution of the latter does not change over time conditional on the treatment and the mediator. We also provide a simulation study and two empirical applications regarding a training programme evaluation and maternity leave reform.

JEL-Codes: C210.

Keywords: direct effects, indirect effects, mediation analysis, changes-in-changes, causal mechanisms, treatment effects.

Martin Huber
Chair of Applied Econometrics – Evaluation of Public Policies
University of Fribourg
Bd. de Pérolles 90
Switzerland – 1700 Fribourg
Martin.Huber@unifr.ch

Mark Schelker
Chair of Public Economics
University of Fribourg
Bd. de Pérolles 90
Switzerland – 1700 Fribourg
Mark.Schelker@unifr.ch

Anthony Strittmatter
Institut Polytechnique de Paris
CREST-ENSAE
5 avenue Henry Le Chatelier
France – 91764 Palaiseau, Cedex
Anthony.Strittmatter@ensae.fr

We have benefited from comments by Giuseppe Germinario as well as conference/seminar participants at the Universities of Neuchâtel, Melbourne, Sydney, Hamburg, and Lisbon, the Luxembourg Institute of Socio-Economic Research, the 2019 meeting of the Austro-Swiss Region of the International Biometric Society in Lausanne, and the 2019 meeting of the International Association for Applied Econometrics in Nicosia.

1 Introduction

Causal mediation analysis aims at disentangling a total treatment effect into an indirect effect operating through an intermediate variable – commonly referred to as mediator – as well as the direct effect. The latter includes any causal mechanisms not operating through the mediator of interest. Even when the treatment is randomly assigned, the mediator is not randomly assigned since it is an intermediate outcome variable. Controlling for the mediator without accounting for its potential endogeneity could introduce selection bias (see, e.g., Robins and Greenland, 1992). Accordingly, controlling for the mediator does generally not identify direct and indirect effects.

This paper suggests a novel identification strategy for causal mediation analysis based on changes-in-changes (CiC) as suggested by Athey and Imbens (2006) for evaluating (total) average and quantile treatment effects. We adapt the approach to the identification of the direct effect and the indirect effect running through a binary mediator. The outcome variable must be continuous and is assumed to be observed both prior to and after treatment and mediator assignment as it is the case in repeated cross sections or panel data. The key identifying assumptions imply that the continuous outcome is strictly monotonic in unobserved heterogeneity and that the distribution of unobserved heterogeneity does not change over time conditional on the treatment and the mediator (the latter assumption is also known as stationarity). Given appropriate common support conditions, this permits identifying direct effects on subpopulations conditional on the treatment and the mediator states, even if both treatment and mediator assignment are endogenous.

Augmenting the assumptions by random treatment assignment and weak monotonicity of the mediator in the treatment allows for causal mediation analysis in subpopulations defined upon whether and how the mediator reacts to the treatment. Specifically, we show the identification of direct effects among those whose mediator is always one (not-affected at 1 in the denomination of Flores and Flores-Lagunes, 2010) and never one (called not-affected at 0) irrespective of treatment

assignment, respectively. Furthermore, we identify the total, direct, and indirect treatment effects on those whose mediator value is affected positively by treatment assignment (called affected positively). For any set of assumptions, we discuss the identification of both average and quantile direct and indirect effects. We note that if appropriately weighted, the respective average effects among not-affected at 0 and 1, and affected positively add up to the average direct and indirect effects in the population.

Identification in the earlier mediation literature typically relied on linear models for the mediator and outcome equations and often neglected endogeneity issues, see for instance Cochran (1957), Judd and Kenny (1981), and Baron and Kenny (1986). More recent contributions use more general identification approaches based on the potential outcome framework and take endogeneity issues explicitly into consideration. Examples include Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), VanderWeele (2009), Imai, Keele, and Yamamoto (2010), Hong (2010), Albert and Nelson (2011), Imai and Yamamoto (2013), Tchetgen Tchetgen and Shpitser (2012), Vansteelandt, Bekaert, and Lange (2012), and Huber (2014). The vast majority of the literature assumes that the covariates observed in the data are sufficiently rich to control for treatment and mediator endogeneity.

Also in empirical economics, there has been an increase in the application of such selection on observables approaches, see for instance Simonsen and Skipper (2006), Flores and Flores-Lagunes (2009), Heckman, Pinto, and Savelyev (2013), Huber (2015), Keele, Tingley, and Yamamoto (2015), Conti, Heckman, and Pinto (2016), Huber, Lechner, and Mellace (2017), Bijwaard and Jones (2019), Bellani and Bia (2018), Huber, Lechner, and Strittmatter (2018), and Doerr and Strittmatter (2019). Comparably few studies in economics develop or apply instrumental variable approaches for disentangling direct and indirect effects, see for instance Powdthavee, Lekfuangfu, and Wooden (2013), Brunello, Fort, Schneeweis, and Winter-Ebmer (2016), Chen, Chen, and Liu (2017), and Frölich and Huber (2017). Our paper

provides another, CiC-based identification strategy that neither rests on selection on observables assumptions nor on instrumental variables for the treatment or the mediator.

While most studies aim at evaluating direct and indirect effects in the total population, a smaller strand of the literature uses the principal stratification framework of Frangakis and Rubin (2002) to investigate effects in subpopulations (or principal strata) defined upon whether and how the mediator reacts to the treatment, see Rubin (2004). This approach has been criticized for typically focussing on direct effects on populations whose mediator is constant (i.e. the not-affected at 0 and 1) rather than decomposing direct and indirect effects for those who are affected and for considering subpopulations rather than the total population, see VanderWeele (2008) and VanderWeele (2012).

Deuchert, Huber, and Schelker (2019) suggest a difference-in-differences (DiD) strategy that alleviates such criticisms. Identification relies on a randomized treatment, monotonicity of the (binary) mediator in the treatment, and particular common trend assumptions on mean potential outcomes across principal strata. The latter imply that mean potential outcomes under specific treatment and mediator states change by the same amount over time across specific subpopulations. Depending on the strength of common trend and effect homogeneity assumptions across principal strata, direct and indirect effects are identified for different subpopulations and under the strongest set of assumptions even for the total population. The approach is based on taking mean differences in observed pre- and post-treatment outcomes within groups defined on treatment and mediator states and appropriately differencing such before-after differences across groups. For instance, under specific common trend assumptions, the direct effect on the not-affected at 0 (never-takers in the denomination of Deuchert, Huber, and Schelker, 2019) is obtained by subtracting the before-after difference in the non-treated and non-mediated group from the before-after difference in the treated and non-mediated group.

Our paper contributes to this literature on principal strata effects, but relies on

different identifying assumptions than Deuchert, Huber, and Schelker (2019). While differential time trends across subpopulations are permitted, our approach restricts the conditional distribution of unobserved heterogeneity over time. The two sets of assumptions are not nested and their appropriateness is to be judged in the empirical context at hand. However, both approaches could be used simultaneously for testing the joint validity of the identifying assumptions of either method, in which case both CiC and DiD converge to the same, true average direct and indirect effects. This may be implemented by a Hausman (1978)-type specification test, e.g. by constructing a t-statistic based on dividing the difference in effects by its standard error (possibly obtained by bootstrapping the difference in effects). As a further distinction to Deuchert, Huber, and Schelker (2019), our method also permits assessing quantile treatment effects (QTEs) rather than average effects only.

In independent work, Sawada (2019) proposes a CiC strategy to tackle non-compliance in randomized experiments when the exclusion restriction of random assignment is violated. While there is an overlap in some identification results of his study and ours (e.g. concerning the direct effect on not-affected at 0), there are also important differences. First, Sawada (2019) predominantly focusses on the average treatment effect on the treated under one-sided non-compliance (ruling out not-affected at 1), which then corresponds to the total effect on those affected positively (compliers in the denomination of Sawada, 2019). Our paper in addition disentangles the total effect on those affected positively into direct and indirect components. Second, under two-sided non-compliance (i.e. the existence of both not-affected at 0 and 1), Sawada (2019) identifies the total effect on those affected positively by assuming homogeneity of the direct effect, while we extend the CiC assumptions to the not-affected at 1 for identifying (direct, indirect, and total) effects on those affected positively as well as the direct effect among those not-affected at 1. Third and in contrast to Sawada (2019), we also provide identification results in the absence of randomization and monotonicity of the mediator in the treatment. On the other hand, Sawada (2019), in contrast to our study, demonstrates that the

CiC strategy does not necessarily require pre-treatment outcomes, but may exploit any pre-treatment variable that has similar rank orders (as a function of unobserved heterogeneity) as the outcome of interest.

As our approach can be used for testing the presence of direct effects and thus, the violation of exclusion restrictions, it is also related to a growing literature on testing identifying assumptions in nonparametric instrumental variable models (when considering the treatment as instrument and the mediator as endogenous treatment). For instance Kitagawa (2015), Huber and Mellace (2015), Mourifié and Wan (2017), Farbmacher and Guber (2018), Sharma (2018), and Wang and Flores-Lagunes (2019) provide tests for moment inequality constraints that are implied by valid instruments and have been derived in Balke and Pearl (1997) and Heckman and Vytlacil (2005). Our paper is also related to studies linking violations of the exclusion restriction to causal mediation analysis, see for instance the partial identification approaches in Flores and Flores-Lagunes (2013) and Chen, Flores, and Flores-Lagunes (2016).

We also provide two empirical applications. The first one reconsiders the Jobs II programme previously analysed by Vinokur, Price, and Schul (1995), a randomized job training intervention designed to analyse the impact of job training on labour market and mental health outcomes. We investigate the direct effect of the randomized offer of treatment on a depression index, as well as its indirect effect through actual participation in the programme as mediator. The reason for investigating the direct effect is that treatment assignment could have a motivation or discouragement effect on those randomly offered or not offered the training. We, however, find the direct effect estimates to be close to zero and statistically insignificant and therefore no indication for the violation of the exclusion restriction when using treatment assignment as instrumental variable for actual participation. In contrast, the moderately negative total and indirect effects on those induced to participate by assignment are statistically significant at least at the 10% level in all but one case and very much in line with the estimate obtained by instrumental variable regression.

In the second application, we investigate the income effect of paid maternity leave

in Switzerland as introduced in 2005. This treatment affects the income of women who become mothers and may thus have an indirect effect operating through the mediator of childbearing. However, the mere availability of paid maternity leave might also affect the outcome through other mechanisms which make up the direct effect, as for instance through a change in statistical discrimination against women (with or without children) by employers. The CiC approach permits evaluating such direct effects among specific groups even under the likely endogeneity of the treatment and the mediator. We find positive direct income effects of roughly 5 percent on treated not affected at 0 (i.e. women aged 20 to 39 without maternal episode even when maternity leave is available) as well as on non-treated (i.e. women aged 46 to 59) for whom paid maternity leave is arguably irrelevant as they are beyond the childbearing age.

The remainder of this study is organized as follows. Section 2 introduces the notation and defines the direct and indirect effects of interest. Section 3 presents the assumptions underlying our CiC approach as well as the identification results. Section 4 provides two applications regarding a training programme evaluation and maternity leave reform. Section 5 concludes. Online Appendices A-D provide the proofs of the identification results. Online Appendix E provides a simulation study in which we compare the CiC to the DiD approach to illustrate our identification results. Online Appendix F provides additional information about the two empirical applications.

2 Notation and effects

2.1 Average effects

Let D denote a binary treatment (e.g., receiving the offer to participate in a training programme) and M a binary intermediate variable or mediator that may be a function of D (e.g., the actual participation in a training programme). Furthermore, let T indicate a particular time period: $T = 0$ denotes the baseline period prior to the

realisation of D and M, $T = 1$ the follow up period after measuring D and M in which the effect of the outcome is evaluated. Finally, let Y_t denote the outcome of interest (e.g., health measures) in period $T = t$. Indexing the outcome by the time period $t \in \{0, 1\}$ implies that it is measured both in the baseline period and after the realisation of D and M. To define the parameters of interest, we make use of the potential outcome notation, see for instance Rubin (1974), and denote by $Y_t(d, m)$ the potential outcome for treatment state $D = d$ and mediator state $M = m$ in time $T = t$, with $d, m, t, \in \{0, 1\}$. Furthermore, let $M(d)$ denote the potential mediator as a function of the treatment state $d \in \{0, 1\}$. For notational ease, we will not use any time index for D and M, because they are assumed to be measured at a single point in time between $T = 0$ and $T = 1$, albeit not necessarily at the same point, as D causally precedes M. Therefore, D and M correspond to the actual treatment and mediator status in $T = 1$, while it is assumed that no treatment or mediation takes place in $T = 0$.

Using this notation, the average treatment effect (ATE) in the ex-post period is defined as $\Delta_1 = E[Y_1(1, M(1)) - Y_1(0, M(0))]$. That is, the ATE corresponds to the effect of D on the outcome that either affects the latter directly (net of any effect on the mediator) or indirectly through an effect on M. Indeed, the total ATE can be disentangled into the direct and indirect effects, denoted by $\theta_1(d) = E[Y_1(1, M(d)) - Y_1(0, M(d))]$ and $\delta_1(d) = E[Y_1(d, M(1)) - Y_1(d, M(0))]$, by adding and subtracting $Y_1(1, M(0))$ or $Y_1(0, M(1))$, respectively:

$$
\begin{aligned}
\Delta_1 &= E[Y_1(1, M(1)) - Y_1(0, M(0))], \\
&= \underbrace{E[Y_1(1, M(1)) - Y_1(1, M(0))]}_{=\delta_1(1)} + \underbrace{E[Y_1(1, M(0)) - Y_1(0, M(0))]}_{=\theta_1(0)}, \\
&= \underbrace{E[Y_1(1, M(1)) - Y_1(0, M(1))]}_{=\theta_1(1)} + \underbrace{E[Y_1(0, M(1)) - Y_1(0, M(0))]}_{=\delta_1(0)}.
\end{aligned}
$$

Distinguishing between $\theta_1(1)$ and $\theta_1(0)$ or $\delta_1(1)$ and $\delta_1(0)$, respectively, implies the possibility of interaction effects between D and M such that the direct and indirect

effects could be heterogeneous across values $d = 1$ and $d = 0$.

In our approach, we consider the concepts of direct and indirect effects within specific subpopulations. The latter are either defined conditional on the treatment and mediator values or conditional on potential mediator values under either treatment states, which matches the so-called principal stratum framework of Frangakis and Rubin (2002). This framework is popular in the instrumental variable literature to stratify the population into never-takers, always-takers, compliers, and defiers based on the potential treatment status as a function of the instrument (e.g. Angrist, Imbens, and Rubin, 1996). In contrast, the mediation literature stratifies the population based on the potential mediator status as a function of the treatment. The instrumental variable approach is nested in the mediation framework when we consider the treatment as instrument and the mediator as treatment. To make the difference to the instrumental variable literature explicit, we adapt in the following the terminology of Flores and Flores-Lagunes (2010).

Any cross-sectional observation unit i (e.g., individual) in the population belongs to one of four strata, henceforth denoted by τ, according to their potential mediator status under either treatment state (see Table 1 for an overview): not-affected at 1 ($n1$: $M(1) = M(0) = 1$) whose mediator is always one (always-takers in the instrumental variable terminology), affected positively (ap: $M(1) = 1$, $M(0) = 0$) whose mediator corresponds to the treatment value (compliers in the instrumental variable terminology), affected negatively (an: $M(1) = 0$, $M(0) = 1$) whose mediator opposes the treatment value (defiers in the instrumental variable terminology), and not affected at 0 ($n0$: $M(1) = M(0) = 0$) whose mediator is never one (never-takers in the instrumental variable terminology). Note that τ cannot be pinned down for any observation unit, because either $M(1)$ or $M(0)$ is observed, but never both. Unless additional assumptions are imposed (such as one-sided non-compliance, e.g., Frölich and Melly, 2013), all strata correspond to latent groups.

Let $\Delta_1^\tau = E[Y_1(1, M(1)) - Y_1(0, M(0))|\tau]$ denote the ATE conditional on $\tau \in \{n1, ap, an, n0\}$; $\theta_1^\tau(d)$ and $\delta_1^\tau(d)$ denote the corresponding direct and indirect effects.

Table 1: Principal strata

$M(1)=1,\ M(0)=1$	Not-affected at 1 (always-takers)	$D=1$	Treated not-affected at 1
		$D=0$	Untreated not-affected at 1
$M(1)=1,\ M(0)=0$	Affected positively (compliers)	$D=1$	Treated affected positively
		$D=0$	Untreated affected positively
$M(1)=0,\ M(0)=1$	Affected negatively (defiers)	$D=1$	Treated affected negatively
		$D=0$	Untreated affected negatively
$M(1)=0,\ M(0)=0$	Not-affected at 0 (never-takers)	$D=1$	Treated not-affected at 0
		$D=0$	Untreated not-affected at 0

Because $M(1) = M(0) = 0$ for any not-affected at 0, the indirect effect for this group is by definition zero ($\delta_1^{n0}(d) = E[Y_1(d,0) - Y_1(d,0)|\tau = n0] = 0$) and $\Delta_1^{n0} = E[Y_1(1,0) - Y_1(0,0)|\tau = n0] = \theta_1^{n0}(1) = \theta_1^{n0}(0) = \theta_1^{n0}$ equals the direct effect for not-affected at 0 (see also the discussion in Section 2.2 of Flores and Flores-Lagunes, 2013). Correspondingly, because $M(1) = M(0) = 1$ for any not-affected at 1, the indirect effect for this group is by definition zero ($\delta_1^{n1}(d) = E[Y_1(d,1) - Y_1(d,1)|\tau = n1] = 0$) and $\Delta_1^{n1} = E[Y_1(1,1) - Y_1(0,1)|\tau = n1] = \theta_1^{n1}(1) = \theta_1^{n1}(0) = \theta_1^{n1}$ equals the direct effect for not-affected at 1. For the affected positively, both direct and indirect effects may exist. Note that $M(d) = d$ due to the definition of affected positively. Accordingly, $\theta_1^{ap}(d) = E[Y_1(1,d) - Y_1(0,d)|\tau = ap]$ equals the direct effect for affected positively, $\delta_1^{ap}(d) = E[Y_1(d,1) - Y_1(d,0)|\tau = ap]$ equals the indirect effect for affected positively, and $\Delta_1^{ap} = E[Y_1(1,1) - Y_1(0,0)|\tau = ap]$ equals the total effect for affected positively. In the absence of any direct effect, the indirect effects on the affected positively are homogeneous, $\delta_1^{ap}(1) = \delta_1^{ap}(0) = \delta_1^{ap} = \Delta_1^{ap}$, and would correspond to the local average treatment effect in the terminology of the instrumental variable approach (LATE, e.g., Angrist, Imbens, and Rubin, 1996). Analogous results hold for the affected negatively.

As already mentioned, we will also consider direct effects conditional on specific values $D = d$ and mediator states $M = M(d) = m$, which are denoted by $\theta_1^{d,m}(d) = E[Y_1(1,m) - Y_1(0,m)|D = d, M(d) = m]$. These parameters are identified under weaker assumptions than strata-specific effects, but are also less straightforward to interpret, as they refer to mixtures of two strata. Instead of stratifying the sample

only by the potential mediator values, it is also possible to stratify the sample by the potential mediator values and the observed treatment status (see last column of Table 1). This leads to eight strata: treated and untreated not-affected at 1, treated and untreated affected positively, treated and untreated affected negatively, as well as treated and untreated not-affected at 0. For instance, we can interpret $\theta_1^{1,0}(1) = E[Y_1(1,0) - Y_1(0,0)|D = 1, M(1) = 0]$ as a mixture of the direct effects for treated not-affected at 0 and treated affected negatively. Likewise, $\theta_1^{0,0}(0)$ refers to untreated not-affected at 0 and untreated affected positively, $\theta_1^{0,1}(0)$ to untreated not-affected at 1 and untreated affected negatively, and $\theta_1^{1,1}(1)$ to treated not-affected at 1 and treated affected positively. The interpretation of the parameters simplifies when the existence of specific strata can be excluded. For example, when 'not-affected at 1' and 'affected negatively' can be ruled-out (one-sided non-compliance), then $\theta_1^{0,0}(0)$ corresponds to the direct effect for the population of all untreated, which is an observable group.

2.2 Quantile effects

We denote by $F_{Y_t(d,m)}(y) = \Pr(Y_t(d,m) \leq y)$ the cumulative distribution function of $Y_t(d,m)$ at outcome level y. Its inverse, $F^{-1}_{Y_t(d,m)}(q) = \inf\{y : F_{Y_t(d,m)}(y) \geq q\}$, is the quantile function of $Y_t(d,m)$ at rank q. The total QTE are denoted by $\Delta_1(q) = F^{-1}_{Y_1(1,M(1))}(q) - F^{-1}_{Y_1(0,M(0))}(q)$. The QTE can be disentangled into the direct quantile effects, denoted by $\theta_1(q,d) = F^{-1}_{Y_1(1,M(d))}(q) - F^{-1}_{Y_1(0,M(d))}(q)$, and the indirect quantile effects, denoted by $\delta_1(q,d) = F^{-1}_{Y_1(d,M(1))}(q) - F^{-1}_{Y_1(d,M(0))}(q)$.

The conditional distribution function in stratum τ is $F_{Y_t(d,m)|\tau}(y) = \Pr(Y_t(d,m) \leq y|\tau)$ and the corresponding conditional quantile function is $F^{-1}_{Y_t(d,m)|\tau}(q) = \inf\{y : F_{Y_t(d,m)|\tau}(y) \geq q\}$ for $\tau \in \{n1, ap, an, n0\}$. Using the previously described stratification framework, we define the QTE conditional on $\tau \in \{n1, ap, an, n0\}$: $\Delta_1^\tau(q) = F^{-1}_{Y_1(1,M(1))|\tau}(q) - F^{-1}_{Y_1(0,M(0))|\tau}(q)$. The direct quantile treatment effect among not-affected at 0 equals $\Delta_1^{n0}(q) = F^{-1}_{Y_1(1,0)|n0}(q) - F^{-1}_{Y_1(0,0)|n0}(q) = \theta_1^{n0}(q)$. The direct quantile effect among not-affected at 1 equals $\Delta_1^{n1}(q) = F^{-1}_{Y_1(1,1)|n1}(q) - F^{-1}_{Y_1(0,1)|n1}(q) =$

$\theta_1^{n1}(q)$. The total QTE among affected positively equals $\Delta_1^{ap}(q) = F^{-1}_{Y_1(1,1)|ap}(q) - F^{-1}_{Y_1(0,0)|ap}(q)$, the direct quantile effect among affected positively equals $\theta_1^{ap}(q,d) = F^{-1}_{Y_1(1,d)|ap}(q) - F^{-1}_{Y_1(0,d)|ap}(q)$, and the indirect quantile effect among affected positively equals $\delta_1^{ap}(q,d) = F^{-1}_{Y_1(d,1)|ap}(q) - F^{-1}_{Y_1(d,0)|ap}(q)$. Finally, we define the direct quantile treatment effects conditional on specific values $D = d$ and mediator states $M = M(d) = m$,

$$\theta_1^{d,m}(q,1) = F^{-1}_{Y_1(1,m)|D=d,M(1)=m}(q) - F^{-1}_{Y_1(0,m)|D=d,M(1)=m}(q) \text{ and}$$
$$\theta_1^{d,m}(q,0) = F^{-1}_{Y_1(1,m)|D=d,M(0)=m}(q) - F^{-1}_{Y_1(0,m)|D=d,M(0)=m}(q),$$

with the quantile function $F^{-1}_{Y_t(d,m)|D=d,M(d)=m}(q) = \inf\{y : F_{Y_t(d,m)|D=d,M(d)=m}(y) \geq q\}$ and the distribution function $F_{Y_t(d,m)|D=d,M(d)=m}(y) = \Pr(Y_t(d,m) \leq y|D = d, M(d) = m)$.

2.3 Observed distribution and quantile transformations

We subsequently define various functions of the observed data required for the identification results. The conditional distribution function of the observed outcome Y_t conditional on treatment value d and mediator state m, is given by $F_{Y_t|D=d,M=m}(y) = \Pr(Y_t \leq y|D = d, M = m)$ for $d, m \in \{0,1\}$. The corresponding conditional quantile function is $F^{-1}_{Y_t|D=d,M=m}(q) = \inf\{y : F_{Y_t|D=d,M=m}(y) \geq q\}$. Furthermore,

$$Q_{dm}(y) := F^{-1}_{Y_1|D=d,M=m} \circ F_{Y_0|D=d,M=m}(y) = F^{-1}_{Y_1|D=d,M=m}(F_{Y_0|D=d,M=m}(y))$$

is the quantile-quantile transform of the conditional outcome from period 0 to 1 given treatment d and mediator status m. This transform maps y at rank q in period 0 ($q = F_{Y_0|D=d,M=m}(y)$) into the corresponding y' at rank q in period 1 ($y' = F^{-1}_{Y_1|D=d,M=m}(q)$).

3 Identification and Estimation

3.1 Identification

This section discusses the identifying assumptions along with the identification results for the various direct and indirect effects. The main identifying assumption is that the unobserved heterogeneity does not change over time conditional on treatment and mediator status. To improve the interpretability of the parameters, we introduce additional independence assumptions about the treatment and weak monotonicity assumptions about the mediator. We note that our assumptions could be adjusted to only hold conditional on a vector of observed covariates. In this case, the identification results would hold within cells defined upon covariate values. In our main discussion, however, covariates are not considered for the sake of ease of notation. For notational convenience, we maintain throughout that $\Pr(T = t, D = d, M = m) > 0$ for $t, d, m \in \{1, 0\}$, implying that all possible treatment-mediator combinations exist in the population in both time periods. Our first assumption implies that potential outcomes are characterized by a continuous nonparametric function, denoted by h, that is strictly monotonic in a scalar U_t that reflects unobserved heterogeneity.

Assumption 1: Strict monotonicity of continuous potential outcomes in unobserved heterogeneity.

The potential outcomes satisfy the following model: $Y_t(d, m) = h(d, m, t, U_t)$, with the general function h being continuous and strictly increasing in the scalar unobservable $U_t \in \mathbb{R}$ for all $d, m, t \in \{0, 1\}$.

Assumption 1 requires the potential outcomes to be continuous implying that there is a one-to-one correspondence between a potential outcome's distribution and quantile functions, which is a condition for point identification. For discrete potential outcomes, only bounds on the effects could be identified, in analogy to the discussion in Athey and Imbens (2006) for total (rather than direct and indirect) effects. Assumption 1 also implies that cross-sectional observation units with identical un-

observed characteristics U_t have the same potential outcomes $Y_t(d,m)$, while higher values of U_t correspond to strictly higher potential outcomes $Y_t(d,m)$. Strict monotonicity is satisfied in additively separable models (e.g., the partial linear model $Y_t(d,m) = g(d,m,t) + U_t$), but Assumption 1 also allows for more flexible non-additive structures that arise in nonparametric models (e.g., the proportional hazard model $Y_t(d,m) = f(d,m)\lambda(t)U_t$).

The next assumption rules out anticipation effects of the treatment or the mediator on the outcome in the baseline period. This assumption is plausible if assignment to the treatment or the mediator cannot be foreseen in the baseline period, such that behavioral changes affecting the pre-treatment outcome are ruled out.

Assumption 2: No anticipation effect of M and D in the baseline period.

$Y_0(d,m) - Y_0(d',m') = 0$, for $d, d', m, m' \in \{1,0\}$.

Similarly, Athey and Imbens (2006) and Chaisemartin and D'Haultfeuille (2018) assume the assignment to the treatment group does not affect the potential outcomes as long as the treatment is not yet realized.

Furthermore, we assume conditional independence between unobserved heterogeneity and time periods conditional on treatment and mediator status.

Assumption 3: Independence of U_t and T conditional on D and M.

(a) $U_t \perp\!\!\!\perp T | D = 1, M = 0$,

(b) $U_t \perp\!\!\!\perp T | D = 0, M = 0$,

(c) $U_t \perp\!\!\!\perp T | D = 0, M = 1$,

(d) $U_t \perp\!\!\!\perp T | D = 1, M = 1$.

For instance, under Assumption 3a, the distribution of U_t is allowed to vary across groups defined upon treatment and mediator state, but not over time within the group with $D = 1, M = 0$. Assumption 3 thus imposes stationarity of U_t within groups defined on D and M. This assumption is weaker than (and thus implied by) requiring that U_t is constant across T for each cross-sectional observation unit i. For example, Assumption 3 is satisfied in the fixed effect model $U_t = \eta + v_t$, with

η being a time-invariant cross-sectional observation unit specific unobservable (fixed effect) and v_t an idiosyncratic time-varying unobservable with the same distribution in both time periods.

Athey and Imbens (2006) and Chaisemartin and D'Haultfeuille (2018) impose time invariance conditional on the treatment status, $U_t \perp\!\!\!\perp T|D = d$, to identify the average treatment effect on the treated, $\Delta_1^{D=1} = E[Y_1(1, M(1)) - Y_1(0, M(0))|D = 1]$ or (using the terminology of instrumental variables) the LATE, $\Delta_1^{pa} = E[Y_1(1, 1) - Y_1(0, 0)|\tau = pa]$, respectively. We additionally condition on the mediator status to identify direct and indirect effects.

For our next assumption, we introduce some further notation. Let $F_{U_t|d,m}(u)) = \Pr(U_t \leq u|D = d, M = m)$ be the conditional distribution of U_t with support $\mathbb{U}_{dm}$.

Assumption 4: Common support.

(a) $\mathbb{U}_{10} \subseteq \mathbb{U}_{00}$, (b) $\mathbb{U}_{00} \subseteq \mathbb{U}_{10}$, (c) $\mathbb{U}_{01} \subseteq \mathbb{U}_{11}$, (d) $\mathbb{U}_{11} \subseteq \mathbb{U}_{01}$.

Assumption 4 is a common support assumption (see discussion in, e.g., Lechner and Strittmatter, 2019). For instance, Assumption 4a implies that any possible value of U_t in the population with $D = 1, M = 0$ is also contained in the population with $D = 0, M = 0$. Assumption 4b imposes that any value of U_t conditional on $D = 0, M = 0$ also exists conditional on $D = 1, M = 0$. Both assumptions together imply that the support of U_t is the same in both populations, albeit the distributions may generally differ. Assumptions 4c and 4d correspond to Assumptions 4a and 4b for $M = 1$ instead of $M = 0$.

Assumptions 1 to 4 permit identifying direct effects on mixed populations, as formally stated in Theorem 1.

Theorem 1: Under Assumptions 1–2,

(a) and Assumptions 3a-3b and 4a, the average and quantile direct effects under $d = 1$ conditional on $D = 1$ and $M(1) = 0$ are identified:

$$\theta_1^{1,0}(1) = E[Y_1 - Q_{00}(Y_0)|D = 1, M = 0],$$
$$\theta_1^{1,0}(q, 1) = F^{-1}_{Y_1|D=1,M=0}(q) - F^{-1}_{Q_{00}(Y_0)|D=1,M=0}(q).$$

(b) and Assumptions 3a-3b and 4b, the average and quantile direct effects under $d = 0$ conditional on $D = 0$ and $M(0) = 0$ are identified:

$$\theta_1^{0,0}(0) = E[Q_{10}(Y_0) - Y_1 | D = 0, M = 0],$$
$$\theta_1^{0,0}(q, 0) = F^{-1}_{Q_{10}(Y_0)|D=0,M=0}(q) - F^{-1}_{Y_1|D=0,M=0}(q).$$

(c) and Assumptions 3c-3d and 4c, the average and quantile direct effects under $d = 0$ conditional on $D = 0$ and $M(0) = 1$ are identified:

$$\theta_1^{0,1}(0) = E[Q_{11}(Y_0) - Y_1 | D = 0, M = 1],$$
$$\theta_1^{0,1}(q, 0) = F^{-1}_{Q_{11}(Y_0)|D=0,M=1}(q) - F^{-1}_{Y_1|D=0,M=1}(q).$$

(d) and Assumptions 3c-3d and 4d, the average and quantile direct effects under $d = 1$ is identified conditional on $D = 1$ and $M(1) = 1$ are identified:

$$\theta_1^{1,1}(1) = E[Y_1 - Q_{01}(Y_0) | D = 1, M = 1],$$
$$\theta_1^{1,1}(q, 1) = F^{-1}_{Y_1|D=1,M=1}(q) - F^{-1}_{Q_{01}(Y_0)|D=1,M=1}(q).$$

Proof. See Online Appendix A.

In the instrumental variable framework, any direct effects of the instrument are typically ruled out by imposing the exclusion restriction, in order to identify the causal effect of an endogenous regressor on the outcome, see for instance Imbens and Angrist (1994). By considering D as instrument and M as endogenous treatment, $\theta_1^{1,0}(1) = \theta_1^{0,0}(0) = \theta_1^{0,1}(0) = \theta_1^{1,1}(1) = 0$ yield testable implications of the exclusion restriction under Assumptions 1-4.

So far, we did not impose exogeneity of the treatment or mediator. In the following, we assume treatment exogeneity by invoking independence between the treatment and the potential post-treatment variables.

Assumption 5: Independence of the treatment and potential mediators/outcomes.

$\{Y_t(d,m), M(d)\} \perp\!\!\!\perp D$, for all $d, m, t, \in \{0,1\}$.

Assumption 5 implies that there are no confounders jointly affecting the treatment on the one hand and the mediator and/or outcome on the other hand. It is satisfied under treatment randomization as in successfully conducted experiments. We could relax Assumption 5, by assuming independence of D conditional on some exogeneous covariates X. However, the implementation in this case would introduces non-trivial practical challenges for the estimation, which we discuss in the next section. Assumption 5 allows identifying the ATE: $\Delta_1 = E[Y_1|D=1] - E[Y_1|D=0]$.

Furthermore, we assume the mediator to be weakly monotonic in the treatment.

Assumption 6: Weak monotonicity of the mediator in the treatment.

$\Pr(M(1) \geq M(0)) = 1$.

Assumption 6 is standard in the instrumental variable literature on LATEs when denoting by D the instrument and by M the endogenous treatment, see Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996), and rules out affected negatively (called defiers in the instrumental variable literature). It is satisfied by design in randomized experiments with one-sided non-compliance, i.e. if no subject randomized out of the treatment ($D = 0$) appears in the subgroup receiving the mediator ($M = 1$), such that affected negatively (defiers) as well not-affected at 1 (always-takers) do not exist. In some contexts, however, the assumption might be disputable. See for instance Angrist and Evans (1998), who use the sex ratio of the first two siblings in a family as instrument for having a third child to estimate the effect of fertility on female labor supply. Monotonicity is motivated by parents' arguable preference for mixed sex siblings in the U.S., inducing affected positively (compliers) to have a third child if the first two are of the same sex. However, monotonicity fails if a subgroup of parents has a preference for at least two children of the same sex and chooses to have a third child if the first two are of mixed sex. That the latter case may be empirically relevant is demonstrated in Lee (2008), who finds that South Korean parents with one son and one daughter are more likely to continue childbearing than parents with two sons.

As discussed in the Online Appendix B, the total ATE $\Delta_1 = E[Y_1|D=1] - E[Y_1|D=0]$ and QTE $\Delta_1(q) = F^{-1}_{Y_1|D=1}(q) - F^{-1}_{Y_1|D=0}(q)$ for the entire population are identified under Assumption 5. Furthermore, Assumptions 5 and 6 yield the strata proportions, denoted by $p_\tau = \Pr(\tau)$, as functions of the conditional mediator probabilities given the treatment, which we denote by $p_{(m|d)} = \Pr(M=m|D=d)$ for $d, m \in \{0,1\}$ (see Online Appendix B):

$$p_{n1} = p_{1|0}, p_{ap} = p_{1|1} - p_{1|0} = p_{0|0} - p_{0|1}, p_{n0} = p_{0|1}. \tag{1}$$

Furthermore, Assumptions 2, 5, and 6 imply that (see Online Appendix B)

$$\Delta_0^{ap} = E[Y_0(1,1) - Y_0(0,0)|ap] = \frac{E[Y_0|D=1] - E[Y_0|D=0]}{p_{1|1} - p_{1|0}} = 0. \tag{2}$$

Therefore, a rejection of the testable implication $E[Y_0|D=1] - E[Y_0|D=0] = 0$ in the data would point to a violation of these assumptions.

Assumptions 5 and 6 permit identifying additional parameters, namely the total, direct, and indirect effects on affected positively, and the direct effects on not-affected at 0 and 1, as shown in Theorems 2 and 3. This follows from the fact that affected negatively are ruled out and that the proportions and potential outcome distributions of the various principal strata are not selective w.r.t. the treatment (which is not excluded in Theorem 1).

Theorem 2: Under Assumptions 1–2, 5-6,

a) and Assumptions 3a-3b and 4a, the average and quantile direct effects on not-affected at 0 are identified:

$$\theta_1^{n0} = \theta_1^{1,0}(1) \text{ and } \theta_1^{n0}(q) = \theta_1^{1,0}(q,1).$$

b) and Assumptions 3a-3b and 4a-4b, the average direct effect under $d=0$ on

affected positively is identified:

$$\theta_1^{ap}(0) = \frac{p_{0|0}}{p_{0|0} - p_{0|1}} \theta_1^{0,0}(0) - \frac{p_{0|1}}{p_{0|0} - p_{0|1}} \theta_1^{1,0}(1).$$

Furthermore, the potential outcome distributions under $d = 0$ on affected positively are identified:

$$\begin{aligned} F_{Y_1(1,0)|ap}(y) = & \frac{p_{0|0}}{p_{0|0} - p_{0|1}} F_{Q_{10}(Y_0)|D=0,M=0}(y) \\ & - \frac{p_{0|1}}{p_{0|0} - p_{0|1}} F_{Y_1|D=1,M=0}(y), \end{aligned} \tag{3}$$

$$\begin{aligned} F_{Y_1(0,0)|ap}(y) = & \frac{p_{0|0}}{p_{0|0} - p_{0|1}} F_{Y_1|D=0,M=0}(y) \\ & - \frac{p_{0|1}}{p_{0|0} - p_{0|1}} F_{Q_{00}(Y_0)|D=1,M=0}(y). \end{aligned} \tag{4}$$

Therefore, the direct quantile effect under $d = 0$ on affected positively, $\theta_1^{ap}(q, 0) = F^{-1}_{Y_1(1,0)|ap}(q) - F^{-1}_{Y_1(0,0)|ap}(q)$, is identified.

c) and Assumptions 3c-3d and 4c, the average and quantile direct effects on not-affected at 1 are identified:

$$\theta_1^{n1} = \theta_1^{0,1}(0) \text{ and } \theta_1^{n1}(q) = \theta_1^{0,1}(q, 0).$$

d) and Assumptions 3c-3d and 4c-4d, the average direct effect under $d = 1$ on affected positively is identified:

$$\theta_1^{ap}(1) = \frac{p_{1|1}}{p_{1|1} - p_{1|0}} \theta_1^{1,1}(1) - \frac{p_{1|0}}{p_{1|1} - p_{1|0}} \theta_1^{0,1}(0).$$

Furthermore, the potential outcome distributions under $d = 1$ for affected positively are identified:

$$\begin{aligned} F_{Y_1(1,1)|ap}(y) = & \frac{p_{1|1}}{p_{1|1} - p_{1|0}} F_{Y_1|D=1,M=1}(y) \\ & - \frac{p_{1|0}}{p_{1|1} - p_{1|0}} F_{Q_{11}(Y_0)|D=0,M=1}(y), \end{aligned} \tag{5}$$

$$F_{Y_1(0,1)|ap}(y) = \frac{p_{1|1}}{p_{1|1}-p_{1|0}} F_{Q_{01}(Y_0)|D=1,M=1}(y) - \frac{p_{1|0}}{p_{1|1}-p_{1|0}} F_{Y_1|D=0,M=1}(y). \tag{6}$$

Therefore, the direct quantile effect under $d=1$ on affected positively $\theta_1^{ap}(q,1) = F^{-1}_{Y_1(1,1)|ap}(q) - F^{-1}_{Y_1(0,1)|ap}(q)$ is identified.

Proof. See Online Appendix C.

Theorem 3: Under Assumptions 1-3 and 5-6,

a) and Assumptions 4a and 4c, the total average treatment effect on affected positively is identified:

$$\begin{aligned}\Delta_1^{ap} =& \frac{p_{1|1}}{p_{1|1}-p_{1|0}} E[Y_1|D=1,M=1] - \frac{p_{1|0}}{p_{1|1}-p_{1|0}} E[Q_{11}(Y_0)|D=0,M=1] \\ & - \frac{p_{0|0}}{p_{1|1}-p_{1|0}} E[Y_1|D=0,M=0] + \frac{p_{0|1}}{p_{1|1}-p_{1|0}} E[Q_{00}(Y_0)|D=1,M=0].\end{aligned}$$

Furthermore, the total quantile treatment effect on affected positively $\Delta_1^{ap}(q) = F^{-1}_{Y_1(1,1)|ap}(q) - F^{-1}_{Y_1(0,0)|ap}(q)$ is identified using the inverse of (5) and (4).

b) and Assumptions 4a and 4d, the average indirect effect under $d=0$ on affected positively is identified:

$$\begin{aligned}\delta_1^{ap}(0) =& \frac{p_{1|1}}{p_{1|1}-p_{1|0}} E[Q_{01}(Y_0)|D=1,M=1] - \frac{p_{1|0}}{p_{1|1}-p_{1|0}} E[Y_1|D=0,M=1] \\ & - \frac{p_{0|0}}{p_{1|1}-p_{1|0}} E[Y_1|D=0,M=0] + \frac{p_{0|1}}{p_{1|1}-p_{1|0}} E[Q_{00}(Y_0)|D=1,M=0].\end{aligned}$$

Furthermore, the quantile indirect effect under $d=0$ on affected positively $\delta_1^{ap}(q,0) = F^{-1}_{Y_1(0,1)|ap}(q) - F^{-1}_{Y_1(0,0)|ap}(q)$ is identified using the inverse of (6) and (4).

c) and Assumptions 4b and 4c, the average indirect effect under $d=1$ on affected

positively is identified:

$$\delta_1^{ap}(1) = \frac{p_{1|1}}{p_{1|1}-p_{1|0}} E[Y_1|D=1,M=1] - \frac{p_{1|0}}{p_{1|1}-p_{1|0}} E[Q_{11}(Y_0)|D=0,M=1]$$
$$- \frac{p_{0|0}}{p_{1|1}-p_{1|0}} E[Q_{10}(Y_0)|D=0,M=0] + \frac{p_{0|1}}{p_{1|1}-p_{1|0}} E[Y_1|D=1,M=0].$$

Furthermore, the quantile indirect effect under $d=1$ on affected positively $\delta_1^{ap}(q,1) = F^{-1}_{Y_1(1,1)|ap}(q) - F^{-1}_{Y_1(1,0)|ap}(q)$ is identified using the inverse of (5) and (3).

Proof. See Online Appendix D.

3.2 Estimation

As in Assumption 5.1 of Athey and Imbens (2006), we assume standard regularity conditions, namely that conditional on $T=t$, $D=d$, and $M=m$, Y is a random draw from that subpopulation defined in terms of $t,d,m \in \{1,0\}$. Furthermore, the outcome in the subpopulations required for the identification results of interest must have compact support and a density that is bounded from above and below as well as continuously differentiable. Denote by N the total sample size across both periods and all treatment-mediator combinations and by $i \in \{1,...,N\}$ an index for the sampled subject, such that (Y_i, D_i, M_i, T_i) correspond to sample realizations of the random variables (Y, D, M, T).

The total, direct, and indirect effects may be estimated using the sample analogy principle, which replaces population moments with sample moments (e.g. Manski, 1988). For instance, any conditional mediator probability given the treatment, $Pr(M=m|D=d)$, is to be replaced by an estimate thereof in the sample, $\frac{\sum_{i=1}^N I\{M_i=m,D_i=d\}}{\sum_{i=1}^N I\{D_i=d\}}$. A crucial step is the estimation of the quantile-quantile transforms. The application of such quantile transformations dates at least back to Juhn, Murphy, and Pierce (1991), see also Chaisemartin and D'Haultfeuille (2018), Wüthrich (2020), and Strittmatter (2019) for recent applications. First, it requires estimating the conditional outcome distribution, $F_{Y_t|D=d,M=m}(y)$, by the conditional

empirical distribution

$$\hat{F}_{Y_t|D=d,M=m}(y) = \frac{1}{\sum_{i=1}^{n} I\{D_i = d, M_i = m, T_i = t\}} \sum_{i:D_i=d,M_i=m,T_i=t} I\{Y_i \leq y\}.$$

Second, inverting the latter yields the empirical quantile function $\hat{F}^{-1}_{Y_t|D=d,M=m}(q)$. The empirical quantile-quantile transform is then obtained by

$$\hat{Q}_{dm}(y) = \hat{F}^{-1}_{Y_1|D=d,M=m}(\hat{F}_{Y_0|D=d,M=m}(y)).$$

This permits estimating the average and quantile effects of interest. Average effects are estimated by replacing any (conditional) expectations with the corresponding sample averages in which the estimated quantile-quantile transforms enter as plug-in estimates. Taking $\theta_1^{1,0}$ (see Theorem 1) as an example, an estimate thereof is

$$\begin{aligned}\hat{\theta}_1^{1,0}(1) =& \frac{1}{\sum_{i=1}^{n} I\{D_i = 1, M_i = 0, T_i = 1\}} \sum_{i:D_i=1,M_i=0,T_i=1} Y_i \\ & - \frac{1}{\sum_{i=1}^{n} I\{D_i = 1, M_i = 0, T_i = 0\}} \sum_{i:D_i=1,M_i=0,T_i=0} \hat{Q}_{00}(Y_i).\end{aligned}$$

Likewise, quantile effects are estimated based on the empirical quantiles.

For the estimation of total ATE and QTE, Athey and Imbens (2006) show that the resulting estimators are $\sqrt{N}$-consistent and asymptotically normal, see their Theorems 5.1 and 5.3. These properties also apply to our context when splitting the sample into subgroups based on the values of a binary treatment and mediator (rather than the treatment only). For instance, the implications of Theorem 1 in Athey and Imbens (2006) when considering subsamples with $D = 1$ and $D = 0$ carry over to considering subsamples with $D = 1, M = 0$ and $D = 0, M = 0$ for estimating the average direct effect on not-affected at 0. In contrast to Athey and Imbens (2006), however, some of our identification results include the conditional mediator probabilities $Pr(M = m|D = d)$. As the latter are estimated with $\sqrt{N}$-consistency, too, it follows that the resulting effect estimators are again $\sqrt{N}$-

consistent and asymptotically normal. We use a non-parametric bootstrap approach to calculate the standard errors. Chaisemartin and D'Haultfeuille (2018) show the validity of the bootstrap approach for such kind of estimators, which follows from their asymptotic normality.

For the case that identifying assumptions to only hold conditional on observed covariates, denoted by X, estimation must be adapted to allow for control variables. Following a suggestion by Athey and Imbens (2006) in their Section 5.1, basing estimation on outcome residuals in which the association of X and Y has been purged by means of a regression is consistent under the additional assumption that the effects of D and M are homogeneous across covariates. As an alternative, Melly and Santangelo (2015) propose a flexible semiparametric estimator that does not impose such a homogeneity-in-covariates assumption and show $\sqrt{N}$-consistency and asymptotic normality.

In Online Appendix E, we show the finite sample performance of the CiC approach. Furthermore, we show the finite sample behavior of the CiC approach under various violations of the identifying assumptions and compare it to the DiD approach of Deuchert, Huber, and Schelker (2019).

4 Applications

4.1 JOBS II Evaluation

Our first empirical application is based on the JOBS II data by Vinokur and Price (1999). JOBS II was a randomized job training intervention in the US, designed to analyse the impact of job training on labour market and mental health outcomes, see Vinokur, Price, and Schul (1995). The JOBS II intervention was conducted in south-eastern Michigan, where 2,464 job seekers were eligible to participate in a randomized field experiment, see Vinokur and Price (1999). We provide further background information about JOBS II in Online Appendix F.1.

We analyse the impact of job training on mental health, namely symptoms of

depression 6 months after training participation. The health outcome (Y) is based on a 11-items index of depression symptoms of the Hopkins Symptom Checklist. For example, respondents were asked how much they were bothered by symptoms such as crying easily, feeling lonely, feeling blue, feeling hopeless, having thoughts of ending their lives, or experiencing a loss of sexual interest. The questions were coded on a 5-point scale, going from 'not at all' (1) to 'extremely' (5), and summarized in a depression variable that consists of the average across all questions.

The study design rules out not-affected at 1 (always-takers) and affected negatively (defiers), because members of the control group did not have access to the job training programme (one-sided non-compliance). 45% of those assigned to training in our data did not participate and are therefore not-affected at 0 (never-takers), the remaining 55% are affected positively (compliers). In order to avoid selection bias w.r.t actual participation, the original JOBS II study by Vinokur, Price, and Schul (1995) analysed the total effect of the policy (i.e. the intention-to-treat effect), including those who, despite receiving an offer to participate, did not take part in the job training. In contrast, we use our methodology to separate the direct effect of mere training assignment, which is our treatment D, from the indirect effect operating through actual training participation, which is our mediator M, among affected positively. We also consider the direct effect on not-affected at 0, which likely differs from that on the affected positively. While being offered (or not offered) the job training might have an effect on the mental health of affected positively by inducing motivation/enthusiasm (or discouragement), it may not have the same effect among not-affected at 0, who do not attend such seminars whatsoever.

More concisely, we base identification on Theorem 2a for the direct effect on not-affected at 0, θ_1^{n0}, on Theorem 2b for the direct effect on affected positively under $d = 0$, $\theta_1^{ap}(0)$, and Theorem 3c for the indirect effect on affected positively under $d = 1$, $\delta_1^{ap}(1)$. None of these approaches requires the presence of not-affected at 1 in the sample. We also note that if random assignment operated through other mechanisms than actual participation in any of the subpopulations as it may appear

Table 2: Descriptive statistics on depression outcomes in pre- and post-mediator periods

	pre-treatment ($T = 0$)		post-mediator ($T = 1$)	
	sample size	mean	sample size	mean
overall	1,796	1.86	1,564	1.73
		(0.58)		(0.67)
$D = 0$	551	1.87	486	1.78
		(0.59)		(0.70)
$D = 1$	1,245	1.86	1,078	1.70
		(0.57)		(0.66)
mean diff		0.01		0.08
pval		0.74		0.03
SD		1.63		11.79

Note: Standard deviations are in parentheses. 'mean diff', 'pval', and 'SD' are the mean difference, its p-value, and the standardized difference, respectively.

reasonable in the context of mental health outcomes, this would violate the exclusion restriction when using assignment as instrumental variable for actual participation in a two stage least squares regression. Given that our identifying assumptions hold, our approach can therefore be used to evaluate the exclusion restriction.

Our evaluation sample consists of a total of 3,360 observations in the pre-treatment and post-mediator periods with non-missing information for D, M, and Y. It is an unbalanced panel due to attrition of roughly 13% of the initial respondents between the two periods. Table 2 provides summary statistics for the outcome in the total sample as well as by treatment group over time. We verify whether randomization was successful by comparing the outcome means of the treatment and control groups in the pre-treatment period ($T = 0$) just prior to the randomization of D. The small difference of 0.01 is not statistically significant according to a two sample t-test. Furthermore, the standardized difference test suggested by Rosenbaum and Rubin (1985) yields a value of just 1.68 and is thus far below 20, a threshold frequently chosen for indicating problematic imbalances across treatment groups. To investigate potential attrition bias we also consider these statistics in the pre-treatment period exclusively among the panel cases that remain in the sample in the post-mediator period (not reported in Table 2). The p-value of the t-test amounts

to 0.52 and the standardized difference of 3.5 is low such that attrition bias does not appear to be a concern. We therefore do not find statistical evidence for a violation of the random assignment of D in our sample. Table 2 also reports the mean difference in outcomes in the post-mediator period ($T = 1$) 6 months after participation, which is an estimate for the total (or intention-to-treat) effect of D. The difference of 0.08 is statistically significant at the 5% level.

Assumption 1 requires our depression measure to be continuous and in a monotonic relationship with unobservables U_t. Essentially, monotonicity with U_t is not testable, however, we see no obvious mechanisms challenging this assumption. To satisfy Assumption 2 (no anticipation), our pre-treatment period is based on data from the screening questionnaire, which makes it impossible to anticipate the outcome of the subsequent randomization of the treatment. As attrition is unlikely to play an essential role, we should be able to rule out one major mechanisms that might challenge Assumptions 3 (independence of U_t and T conditional on D and M) and 4 (common support). Again, we see no obvious further challenges. Assumption 5 (independence of the treatment and potential mediators/outcomes) holds due to the arguably successful randomization. Assumption 6 (no affected negatively) is met at least under $d = 0$, as participation is not possible without eligibility. Moreover, it appears difficult to argue that an individual would avoid the job training just because it has been randomly chosen to being eligible.

Table 3: Empirical results for Jobs II

	Changes-in-Changes				Difference-in-Differences				Type shares	
	$\hat{\theta}_1^{n0}$	$\hat{\Delta}_1^{ap}$	$\hat{\theta}_1^{ap}(0)$	$\hat{\delta}_1^{ap}(1)$	$\hat{\theta}_1^{n0}$	$\hat{\Delta}_1^{ap}$	$\hat{\theta}_1^{ap}(0)$	$\hat{\delta}_1^{ap}(1)$	$\hat{p}_{n0}$	$\hat{p}_{pa}$
est	-0.04	-0.11	0.06	-0.17	-0.03	-0.12	-0.06	-0.06	0.45	0.55
se	0.05	0.06	0.05	0.08	0.05	0.06	0.05	0.07	0.01	0.01
pval	0.40	0.06	0.26	0.04	0.52	0.03	0.21	0.43	0.00	0.00

Note: 'est', 'se', and 'pval' provide the effect estimate, standard error, and p-value of the respective estimator. $\hat{p}(n)$ and $\hat{p}(c)$ are the estimated never-taker and complier shares. Standard errors are based on cluster bootstrapping the effects 1999 times where clustering is on the respondent level.

Table 3 presents the estimation results based on our CiC approach and the DiD strategy of Deuchert, Huber, and Schelker (2019) when (linearly) controlling for the

gender of respondents in either case. Standard errors rely on cluster bootstrapping the direct and indirect effects 1999 times, where clustering is on the respondent level. The CiC and DiD estimates of the direct effects on not-affected at 0, $\hat{\theta}_1^{n0}(0)$, as well as on affetced positively, $\hat{\theta}_1^{ap}(0)$, are not statistically significant at conventional levels. Hence, we do not find statistical evidence for a direct effect of the mere assignment into the training programme on the depression outcome. In an instrumental variable setup, such a relationship would point to a violation of the exclusion restriction when using assignment as instrument for participation. In contrast, we find for both CiC and DiD negative total effects among affected positively $\hat{\Delta}_1^{ap}$ that are statistically significant at least at the 10% level. In the case of CiC, also the negative indirect effect among affected positively, $\hat{\delta}_1^{ap}(1)$, is significant at the 5% level, while this is not the case for DiD. By and large, our results point to a moderately negative treatment effect on depressive symptoms through actual programme participation, rather than through other (i.e. direct) mechanisms. In comparison to an IV approach, we note that the CiC estimates $\hat{\delta}_1^{ap}(1)$ and $\hat{\Delta}_1^{ap}$ as well as the DiD estimate $\hat{\Delta}_1^{ap}$ are in fact rather similar to the result of a two stage least squares regression relying on the exclusion restriction by using D as instrument for M. The latter approach yields a LATE in the post-mediator period of -0.14 with a heteroskedasticity-robust standard error of 0.07 (significant at the 5% level).

4.2 Paid Maternity Leave Reform

In many empirical problems, treatment randomization is violated. Theorem 1 provides conditions under which direct effects for specific subpopulations are identified even in the absence of this assumption. To illustrate this approach, we analyse the effect of paid maternity leave on labour income as introduced in July 2005 in Switzerland (Online Appendix F.2 provides background information about the reform), using data from the Swiss Labour Force Survey of the Federal Statistical Office (FSO). The FSO interviews roughly 50,000 households in a rotating 5-year panel and we consider information from the waves of 2004, 2006 and 2007.

Our pre-treatment period consists of observations from the first six month of 2004, well before the federal referendum on paid maternity leave in late September 2004, in order to avoid anticipation effects (Assumption 2). Political campaigning and discussions about a referendum typically start two to three months earlier. Given that all previous attempts to introduce paid maternity leave were rejected in popular ballots, the latest in 1999, and that the subsequent acceptance with 55.4% was far from overwhelming, important anticipation effects appear unlikely.

The treatment is defined as the option to take paid maternity leave in case of giving birth, which is only relevant for fertile women. For this reason, we define the age group between 20 and 39 years as treatment group. Females in the age group between 46 and 59 years are considered as control group as they are arguably beyond the childbearing age. The mediator is defined as maternal episode in the waves of 2006 and 2007. We do not use data around the introduction of paid maternity leave legislation in 2005. Excluding mothers with a maternal episode already in the pre-treatment period (2004) from our sample, we obtain a sample of 4627 females.

The outcome measures are gross and net income in 2007. Conditional independence between time and unobserved heterogeneity (Assumption 3) and common support (Assumption 4) assumptions appear plausible as there were no further interventions or specific turmoil in Swiss labour markets during this period. The groups 'not-affected at 1' and 'affected negatively' do not exist in our sample, because we observe no maternal episodes in the control group. Therefore, weak monotonicity (Assumption 6) holds in our sample, even though it could be violated in principle if females past 45 years were to give birth.

We cannot plausibly estimate direct and indirect effects based on Theorems 2 and 3, as Assumptions 5 (independence of the treatment and potential mediators/outcomes) is unlikely to hold. The option to take paid maternity leave is not randomly assigned, but depends on age and is most likely associated with income. However, Theorem 1 permits estimating the direct effects of paid leave, which might operate through general equilibrium effects, on 'treated not-affected at 0' ($\theta_1^{1,0}(1)$

based on Theorem 1a) as well as on 'non-treated not-affected at 0 and non-treated affected positively' ($\theta_1^{0,0}(0)$ based on Theorem 1b). Both groups bear a clear interpretation.

The group 'treated not-affected at 0' are women aged between 20 to 39 years who do not give birth, despite having the option to receive paid maternity leave. Older women could also be not-affected at 0, but they are not treated, because they are arguably beyond the childbearing age. We explicitly allow for the possibility that younger women are affected positively, but we cannot identify the effects for this group.

The group of 'non-treated not-affected at 0 and non-treated affected positively' contains all women aged between 46 to 59 years. We argued above that all women aged between 46 to 59 years are non-treated, because the reform does not give them the option to take paid maternity leave (with high probability). Furthermore, we argued above that the groups 'affected negatively' and 'not-affected at 1' do not exist. Accordingly, the groups of 'non-treated not-affected at 0' and 'non-treated affected positively' correspond to the observable group of all non-treated.

Table 4: Empirical results for introduction of maternity leave in Switzerland

	Gross income		Net income	
	$\theta_1^{1,0}(1)$	$\theta_1^{0,0}(0)$	$\theta_1^{1,0}(1)$	$\theta_1^{0,0}(0)$
est	2436.66	2665.33	2122.68	2259.58
se	1417.73	1561.58	1252.53	1348.18
pval	0.09	0.09	0.09	0.09

Note: 'est', 'se', and 'pval' provide the effect estimate, standard error, and p-value of the respective estimator. $\theta_1^{1,0}(1)$ is the direct effect on treated never-takers and $\theta_1^{0,0}(0)$ is the direct effect on non-treated compliers and never takers. Standard errors are based on cluster bootstrapping the effects 1999 times where clustering is on the individual level.

The results in Table 4 suggest positive direct effects on treated not-affected at 0 (younger females without maternal episode) and on the non-treated (older females). The estimates for treated not-affected at 0 amount to CHF 2437 in gross income or CHF 2123 in net income, which corresponds to an increase of about 4.8% relative to the respective average. The effects on the non-treated are slightly larger and amount

to a CHF 2665 increase in gross income (5.2%) and CHF 2259 in net income (5.0%). All point estimates are marginally significant at the 10% level when clustering on the individual level. The introduction of paid maternity leave arguably reduced the financial burden of maternity on firms and affected households, thus, making female workers relatively less expensive in expectation. From an economic perspective, this likely triggered adjustments in the labour markets and, through general equilibrium effects, also affected groups that did not take paid leave. The direct effect on younger females without a maternal episode, for instance, could point towards a reduction in statistical discrimination by employers.

5 Conclusion

We proposed a novel identification strategy for causal mediation analysis with repeated cross sections or panel data based on changes-in-changes (CiC) assumptions that are related but yet different to Athey and Imbens (2006) considering total treatment effects. Strict monotonicity of outcomes in unobserved heterogeneity and distributional time invariance of the latter within groups defined on treatment and mediator states are key assumptions for identifying direct effects within these groups. Additionally assuming random treatment assignment and weak monotonicity of the mediator in the treatment permits identifying direct effects on not-affected at 0 and 1 (never- and always-takers in the instrumental variable terminology) as well as total, direct, and indirect effects on affected positively (compliers in the instrumental variable terminology). We also provided two empirical applications to the Jobs II programme and the introduction of paid maternity leave in Switzerland.

'Not-affected at 0 and 1' as well as 'affected positively' are latent groups. Without random treatment assignment, we can only identify effects for mixtures between two latent groups conditional on the treatment status. This complicates the interpretation of the results, unless some latent groups can be excluded because of the empirical design (see, e.g., the discussion about one-sided non-compliance in Frölich and Melly, 2013).

References

Albert, J. M., and S. Nelson (2011): "Generalized Causal Mediation Analysis," *Biometrics*, 67, 1028–1038.

Angrist, J., and W. Evans (1998): "Children and their Parents Labor Supply: Evidence from Exogeneous Variation in Family Size," *American Economic Review*, 88, 450–477.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–472.

Athey, S., and G. W. Imbens (2006): "Identification and Inference in Nonlinear Difference-In-Difference Models," *Econometrica*, 74, 431–497.

Balke, A., and J. Pearl (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

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, 1173–1182.

Bellani, L., and M. Bia (2018): "The Long-Run Effect of Childhood Poverty and the Mediating Role of Education," *Journal of the Royal Statistical Society: Series A*, 182, 37–68.

Bijwaard, G. E., and A. M. Jones (2019): "An IPW Estimator for Mediation Effects in Hazard Models: With an Application to Schooling, Cognitive Ability and Mortality," *Empirical Economics*, 57, 129–175.

Brunello, G., M. Fort, N. Schneeweis, and R. Winter-Ebmer (2016): "The Causal Effect of Education on Health: What is the Role of Health Behaviors?," *Health Economics*, 25, 314–336.

Chaisemartin, C., and X. D'Haultfeuille (2018): "Fuzzy Differences-in-Differences," *Review of Economic Studies*, 85, 999–1028.

Chen, S. H., Y. C. Chen, and J. T. Liu (2017): "The Impact of Family Composition on Educational Achievement," *Journal of Human Resources*, forthcoming.

Chen, X., C. Flores, and A. Flores-Lagunes (2016): "Bounds on Average Treatment Effects with an Invalid Instrument: An Application to the Oregon Health Insurance Experiment," *Working Paper*.

Cochran, W. G. (1957): "Analysis of Covariance: Its Nature and Uses," *Biometrics*, 13, 261–281.

Conti, G., J. J. Heckman, and R. Pinto (2016): "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour," *Economic Journal*, 126, F28–F65.

Deuchert, E., M. Huber, and M. Schelker (2019): "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*, 37, 710–720.

Doerr, A., and A. Strittmatter (2019): "Identifying Causal Channels of Policy Reforms with Multiple Treatments and Different Types of Selection," *Working Paper*.

Farbmacher, H., and R. Guber (2018): "Instrument Validity Tests with Causal Trees," *Working Paper*.

Flores, C. A., and A. Flores-Lagunes (2009): "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness," *IZA Discussion Paper No. 4237*.

Flores, C. A., and A. Flores-Lagunes (2010): "Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects," *mimeo, University of Florida*.

——— (2013): "Partial Identification of Local Average Treatment Effects with an Invalid Instrument," *Journal of Business & Economic Statistics*, 31, 534–545.

Frangakis, C., and D. Rubin (2002): "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

Frölich, M., and B. Melly (2013): "Identification of treatment effects on the treated with one-sided non-compliance," *Econometric Reviews*, 32(3), 384–414.

Frölich, M., and M. Huber (2017): "Direct and Indirect Treatment Effects – Causal Chains and Mediation Analysis with Instrumental Variables," *Journal of the Royal Statistical Society: Series B*, 79, 1645–1666.

Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–71.

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, 2052–2086.

Heckman, J. J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

Hong, G. (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.

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

Huber, M. (2015): "Causal Pitfalls in the Decomposition of Wage Gaps," *Journal of Business & Economic Statistics*, 33, 179–191.

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

Huber, M., M. Lechner, and A. Strittmatter (2018): "Direct and Indirect Effects of Training Vouchers for the Unemployed," *Journal of the Royal Statistical Society: Series A*, 181, 441–463.

Huber, M., and G. Mellace (2015): "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, 97, 398–411.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Imai, K., and T. Yamamoto (2013): "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments," *Political Analysis*, 21, 141–171.

Imbens, G. W., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Judd, C. M., and D. A. Kenny (1981): "Process Analysis: Estimating Mediation in Treatment Evaluations," *Evaluation Review*, 5, 602–619.

Juhn, C., K. M. Murphy, and B. Pierce (1991): "Accounting for the Slowdown in Black-White Wage Convergence," in *Workers and their Wages: Changing Patterns in the United States*, ed. by M. H. Kosters, pp. 107–143. American Enterprise Insitute Predd, Washington, DC.

Keele, L., D. Tingley, and T. Yamamoto (2015): "Identifying Mechanisms

Behind Policy Interventions via Causal Mediation Analysis," *Journal of Policy Analysis and Management*, 34, 937–963.

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063.

Lechner, M., and A. Strittmatter (2019): "Practical Procedures to Deal with Common Support Problems in Matching Estimation," *Econometric Reviews*, 38, 193–207.

Lee, J. (2008): "Sibling Size and Investment in Children's Education: An Asian Instrument," *Journal of Population Economics*, 21, 855–875.

Manski, C. F. (1988): *Analog Estimation Methods in Econometrics*. Chapman & Hall, New York.

Melly, B., and G. Santangelo (2015): "The Changes-in-Changes Model with Covariates," *Working Paper*.

Mourifié, I., and Y. Wan (2017): "Testing LATE Assumptions," *Review of Economics and Statistics*, 99, 305–313.

Pearl, J. (2001): "Direct and Indirect Effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco. Morgan Kaufman.

Petersen, M. L., S. E. Sinisi, and M. J. van der Laan (2006): "Estimation of Direct Causal Effects," *Epidemiology*, 17, 276–284.

Powdthavee, N., W. N. Lekfuangfu, and M. Wooden (2013): "The Marginal Income Effect of Education on Happiness: Estimating the Direct and Indirect Effects of Compulsory Schooling on Well-Being in Australia," *IZA Discussion Paper No. 7365*.

Robins, J. M. (2003): "Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects," in *Highly Structured Stochastic Systems*, ed. by P. Green, N. Hjort, and S. Richardson, pp. 70–81. Oxford University Press, UK.

Robins, J. M., and S. Greenland (1992): "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143–155.

Rosenbaum, P. R., and D. B. Rubin (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *American Statistician*, 39, 33–38.

Rubin, D. B. (1974): "Estimating the Causal Effect of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Rubin, D. B. (2004): "Direct and Indirect Causal Effects via Potential Outcomes," *Scandinavian Journal of Statistics*, 31, 161–170.

Sawada, M. (2019): "Non-Compliance in Randomized Control Trials without Exclusion Restrictions," *Working Paper*.

Sharma, A. (2018): "Necessary and Probably Sufficient Test for Finding Valid Instrumental Variables," *arXiv:1812.01412*.

Simonsen, M., and L. Skipper (2006): "The Costs of Motherhood: An Analysis Using Matching Estimators," *Journal of Applied Econometrics*, 21, 919–934.

Strittmatter, A. (2019): "Heterogeneous Earnings Effects of the Job Corps by Gender: A Translated Quantile Approach," *Labour Economics*, 61, 1–13.

Tchetgen Tchetgen, E. J., and I. Shpitser (2012): "Semiparametric Theory for Causal Mediation Analysis: Efficiency Bounds, Multiple Robustness, and Sensitivity Analysis," *Annals of Statistics*, 40, 1816–1845.

VanderWeele, T. J. (2008): "Simple Relations Between Principal Stratification and Direct and Indirect Effects," *Statistics & Probability Letters*, 78, 2957–2962.

VanderWeele, T. J. (2009): "Marginal Structural Models for the Estimation of Direct and Indirect Effects," *Epidemiology*, 20, 18–26.

VanderWeele, T. J. (2012): "Comments: Should Principal Stratification Be Used to Study Mediational Processes?," *Journal of Research on Educational Effectiveness*, 5, 245–249.

Vansteelandt, S., M. Bekaert, and T. Lange (2012): "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects," *Epidemiologic Methods*, 1, 129–158.

Vinokur, A. D., and R. H. Price (1999): "Jobs II Preventive Intervention for Unemployed Job Seekers, 1991-1993," *Inter-University Consortium for Political and Social Research*.

Vinokur, A. D., R. H. Price, and Y. Schul (1995): "Impact of the JOBS Intervention on Unemployed Workers Varying in Risk for Depression," *American Journal of Community Psychology*, 23, 39–74.

Wang, X., and A. Flores-Lagunes (2019): "Conscription and Military Service: Do They Result in Future Violent and Non-Violent Incarcerations and Recidivism?," *Working Paper*.

Wüthrich, K. (2020): "A Comparison of Two Quantile Models with Endogeneity," *Journal of Business & Economic Statistics*, 38, 443–456.

Online Appendices to "Direct and Indirect Effects based on Changes-in-Changes"

Martin Huber[†], Mark Schelker[†], Anthony Strittmatter[‡]

[†]University of Fribourg, Dept. of Economics

[‡]CREST-ENSAE, Institut Polytechnique (IP) Paris

Abstract: We propose a novel approach for causal mediation analysis based on changes-in-changes assumptions restricting unobserved heterogeneity over time. This allows disentangling the causal effect of a binary treatment on a continuous outcome into an indirect effect operating through a binary intermediate variable (called mediator) and a direct effect running via other causal mechanisms. We identify average and quantile direct and indirect effects for various subgroups under the condition that the outcome is monotonic in the unobserved heterogeneity and that the distribution of the latter does not change over time conditional on the treatment and the mediator. We also provide a simulation study and two empirical applications regarding a training programme evaluation and maternity leave reform.

Keywords: Direct effects, indirect effects, mediation analysis, changes-in-changes, causal mechanisms, treatment effects.

JEL classification: C21.

Sections:

A Proof of Theorem 1

A.1 Average direct effect under d = 1 conditional on D = 1 and M(1) = 0

In the following, we prove that $\theta_1^{1,0}(1) = E[Y_1(1,0) - Y_1(0,0)|D = 1, M_i(1) = 0] = E[Y_1 - Q_{00}(Y_0)|D = 1, M = 0]$. Using the observational rule, we obtain $E[Y_1(1,0)|D = 1, M(1) = 0] = E[Y_1|D = 1, M = 0]$. Accordingly, we have to show that $E[Y_1(0,0)|D = 1, M(1) = 0] = E[Q_{00}(Y_0)|D = 1, M = 0]$ to finish the proof.

Denote the inverse of $h(d, m, t, u)$ by $h^{-1}(d, m, t; y)$, which exists because of the strict monotonicity required in Assumption 1. Under Assumptions 1 and 3a, the conditional potential outcome distribution function equals

$$\begin{aligned} F_{Y_t(d,0)|D=1,M=0}(y) &\overset{A1}{=} \Pr(h(d,m,t,U_t) \leq y|D=1, M=0, T=t), \\ &= \Pr(U_t \leq h^{-1}(d,m,t;y)|D=1, M=0, T=t), \\ &\overset{A3a}{=} \Pr(U_t \leq h^{-1}(d,m,t;y)|D=1, M=0), \\ &\overset{A3a}{=} \Pr(U_{t'} \leq h^{-1}(d,m,t;y)|D=1, M=0), \\ &= F_{U_{t'}|10}(h^{-1}(d,m,t;y)), \end{aligned} \tag{A.1}$$

for $d, t, t' \in \{0, 1\}$. We use these quantities in the following.

First, evaluating $F_{Y_1(0,0)|D=1,M=0}(y)$ at $h(0,0,1,u)$ gives

$$F_{Y_1(0,0)|D=1,M=0}(h(0,0,1,u)) = F_{U_t|10}(h^{-1}(0,0,1;h(0,0,1,u))) = F_{U_t|10}(u),$$

for any $t \in \{0,1\}$. Applying $F^{-1}_{Y_1(0,0)|D=1,M=0}(q)$ to both sides, we have

$$h(0,0,1,u) = F^{-1}_{Y_1(0,0)|D=1,M=0}(F_{U_t|10}(u)). \tag{A.2}$$

Second, for $F_{Y_0(0,0)|D=1,M=0}(y)$ we have

$$F^{-1}_{U_t|10}(F_{Y_0(0,0)|D=1,M=0}(y)) = h^{-1}(0,0,0;y). \tag{A.3}$$

Combining (A.2) and (A.3) yields,

$$h(0,0,1,h^{-1}(0,0,0;y)) = F^{-1}_{Y_1(0,0)|D=1,M=0} \circ F_{Y_0(0,0)|D=1,M=0}(y). \tag{A.4}$$

Note that $h(0,0,1,h^{-1}(0,0,0;y))$ maps the period 1 (potential) outcome of an cross-sectional observation unit with the outcome y in period 0 under non-treatment without the mediator. Accordingly, $E[F^{-1}_{Y_1(0,0)|D=1,M=0} \circ F_{Y_0(0,0)|D=1,M=0}(Y_0)|D=1,M=0] = E[Y_1(0,0)|D=1,M=0]$. We can identify $F_{Y_0(0,0)|D=1,M=0}(y)$ under Assumption 2, but we cannot identify $F_{Y_1(0,0)|D=1,M=0}(y)$. However, we show in the following that we can identify the overall quantile-quantile transform $F^{-1}_{Y_1(0,0)|D=1,M=0} \circ F_{Y_0(0,0)|D=1,M=0}(y)$ under the additional Assumption 3b.

Under Assumptions 1 and 3b, the conditional potential outcome distribution function equals

$$\begin{aligned}
F_{Y_t(d,0)|D=0,M=0}(y) &\overset{A1}{=} \Pr(h(d,m,t,U_t) \leq y | D=0,M=0,T=t), \\
&= \Pr(U_t \leq h^{-1}(d,m,t;y) | D=0,M=0,T=t), \\
&\overset{A3b}{=} \Pr(U_t \leq h^{-1}(d,m,t;y) | D=0,M=0), \\
&\overset{A3b}{=} \Pr(U_{t'} \leq h^{-1}(d,m,t;y) | D=0,M=0), \\
&= F_{U_{t'}|00}(h^{-1}(d,m,t;y)),
\end{aligned} \tag{A.5}$$

for $d, t, t' \in \{0,1\}$. We repeat similar steps as above. First, evaluating $F_{Y_1(0,0)|D=0,M=0}(y)$ at $h(0,0,1,u)$ gives

$$F_{Y_1(0,0)D=0,M=0}(h(0,0,1,u)) = F_{U_t|00}(h^{-1}(0,0,1;h(0,0,1,u))) = F_{U_t|00}(u),$$

for any $t \in \{0,1\}$. Applying $F^{-1}_{Y_1(0,0)|D=0,M=0}(q)$ to both sides, we have

$$h(0,0,1,u) = F^{-1}_{Y_1(0,0)|D=0,M=0}(F_{U_t|00}(u)). \tag{A.6}$$

Second, for $F_{Y_0(0,0)|D=0,M=0}(y)$ we have

$$F^{-1}_{U_t|00}(F_{Y_0(0,0)|D=0,M=0}(y)) = h^{-1}(0,0,0;y). \tag{A.7}$$

Combining (A.6) and (A.7) yields,

$$h(0,0,1,h^{-1}(0,0,0;y)) = F^{-1}_{Y_1(0,0)|D=0,M=0} \circ F_{Y_0(0,0)|D=0,M=0}(y). \tag{A.8}$$

The left sides of (A.4) and (A.8) are equal. In contrast to (A.4), (A.8) contains only distributions that can be identified from observable data. In particular, $F_{Y_t(0,0)|D=0,M=0}(y) = \Pr(Y_t(0,0) \leq y|D=0,M=0) = \Pr(Y_t \leq y|D=0,M=0)$. Accordingly, we can identify $F^{-1}_{Y_1(0,0)|D=1,M=0} \circ F_{Y_0(0,0)|D=1,M=0}(y)$ by $Q_{00}(y) \equiv F^{-1}_{Y_1|D=0,M=0} \circ F_{Y_0|D=0,M=0}(y)$.

Parsing Y_0 through $Q_{00}(\cdot)$ in the treated group without mediator gives

$$\begin{aligned}
&E[Q_{00}(Y_0)|D=1,M=0] \\
&\quad = E[F^{-1}_{Y_1|D=0,M=0} \circ F_{Y_0|D=0,M=0}(Y_0)|D=1,M=0], \\
&\quad = E[F^{-1}_{Y_1(0,0)|D=0,M=0} \circ F_{Y_0(0,0)|D=0,M=0}(Y_0(1,0))|D=1,M=0], \\
&\quad \overset{A1,A3b}{=} E[h(0,0,1,h^{-1}(0,0,0;Y_0(1,0)))|D=1,M=0], \\
&\quad \overset{A2}{=} E[h(0,0,1,h^{-1}(0,0,0;Y_0(0,0)))|D=1,M=0], \\
&\quad \overset{A1,A3a}{=} E[F^{-1}_{Y_1(0,0)|D=1,M=0} \circ F_{Y_0(0,0)|D=1,M=0}(Y_0(0,0))|D=1,M=0], \\
&\quad = E[Y_1(0,0)|D=1,M=0] = E[Y_1(0,0)|D=1,M(1)=0],
\end{aligned} \tag{A.9}$$

which has data support because of Assumption 4a.

A.2 Quantile direct effect under d = 1 conditional on D = 1 and M(1) = 0

In the following, we prove that

$$\begin{aligned}\theta_1^{1,0}(q,1) &= F^{-1}_{Y_1(1,0)|D=1,M(1)=0}(q) - F^{-1}_{Y_1(0,0)|D=1,M(1)=0}(q),\\ &= F^{-1}_{Y_1|D=1,M=0}(q) - F^{-1}_{Q_{00}(Y_0)|D=1,M=0}(q).\end{aligned}$$

For this purpose, we have to show that

$$F_{Y_1(1,0)|D=1,M(1)=0}(y) = F_{Y_1|D=1,M=0}(y) \text{ and} \tag{A.10}$$

$$F_{Y_1(0,0)|D=1,M(1)=0}(y) = F_{Q_{00}(Y_0)|D=1,M=0}(y), \tag{A.11}$$

which is sufficient to show that the quantiles are also identified. We can show (A.10) using the observational rule $F_{Y_1(1,0)|D=1,M(1)=0}(y) = F_{Y_1|D=1,M=0}(y) = E[1\{Y_1 \leq y\}|D=1,M=0]$, with $1\{\cdot\}$ being the indicator function.

In analogy to (A.9), we obtain

$$\begin{aligned}&F_{Q_{00}(Y_0)|D=1,M=0}(y)\\ &\quad= E[1\{Q_{00}(Y_0) \leq y\}|D=1,M=0],\\ &\quad= E[1\{F^{-1}_{Y_1|D=0,M=0} \circ F_{Y_0|D=0,M=0}(Y_0) \leq y\}|D=1,M=0],\\ &\quad= E[1\{Y_1(0,0) \leq y\}|D=1,M=0],\\ &\quad= F_{Y_1(0,0)|D=1,M(1)=0}(y),\end{aligned} \tag{A.12}$$

which proves (A.11).

A.3 Average direct effect under d = 0 conditional on D = 0 and M(0) = 0

In the following, we show that $\theta_1^{0,0}(0) = E[Y_1(1,0) - Y_1(0,0)|D = 0, M(0) = 0] = E[Q_{10}(Y_0) - Y_1|D = 0, M = 0]$. Using the observational rule, we obtain $E[Y_1(0,0)|D = 0, M(0) = 0] = E[Y_1|D = 0, M = 0]$. Accordingly, we have to show that $E[Y_1(1,0)|D = 0, M(0) = 0] = E[Q_{10}(Y_0)|D = 0, M = 0]$ to finish the proof.

First, we use (A.5) to evaluate $F_{Y_1(1,0)|D=0,M=0}(y)$ at $h(1,0,1,u)$

$$F_{Y_1(1,0)|D=0,M=0}(h(1,0,1,u)) = F_{U_t|10}(h^{-1}(1,0,1;h(1,0,1,u))) = F_{U_t|10}(u),$$

for any $t \in \{0,1\}$. Applying $F^{-1}_{Y_1(1,0)|D=0,M=0}(q)$ to both sides, we have

$$h(1,0,1,u) = F^{-1}_{Y_1(1,0)|D=0,M=0}(F_{U_t|10}(u)). \tag{A.13}$$

Second, for $F_{Y_0(1,0)|D=0,M=0}(y)$ we have

$$F^{-1}_{U_t|10}(F_{Y_0(1,0)|D=0,M=0}(y)) = h^{-1}(1,0,0;y), \tag{A.14}$$

using (A.5). Combining (A.13) and (A.14) yields,

$$h(1,0,1,h^{-1}(1,0,0;y)) = F^{-1}_{Y_1(1,0)|D=0,M=0} \circ F_{Y_0(1,0)|D=0,M=0}(y). \tag{A.15}$$

Note that $h(1,0,1,h^{-1}(1,0,0;y))$ maps the period 1 (potential) outcome of an cross-sectional observation unit with the outcome y in period 0 under treatment without the mediator. Accordingly, $E[F^{-1}_{Y_1(1,0)|D=0,M=0} \circ F_{Y_0(1,0)|D=0,M=0}(Y_0)|D = 0, M = 0] = E[Y_1(1,0)|D = 1, M = 0]$. We can identify $F_{Y_0(1,0)|D=0,M=0}(y)$ under Assumption 2, but we cannot identify $F_{Y_1(1,0)|D=0,M=0}(y)$. However, we show in the following that we can identify the overall quantile-quantile transform $F^{-1}_{Y_1(1,0)|D=0,M=0} \circ F_{Y_0(1,0)|D=0,M=0}(y)$ under the additional Assumption 3a.

First, we use (A.1) to evaluate $F_{Y_1(1,0)|D=1,M=0}(y)$ at $h(1,0,1,u)$

$$F_{Y_1(1,0)|D=10,M=0}(h(1,0,1,u)) = F_{U_t|10}(h^{-1}(1,0,1;h(1,0,1,u))) = F_{U_t|10}(u),$$

for any $t \in \{0,1\}$. Applying $F^{-1}_{Y_1(1,0)|D=1,M=0}(q)$ to both sides, we have

$$h(1,0,1,u) = F^{-1}_{Y_1(1,0)|D=1,M=0}(F_{U_t|10}(u)). \tag{A.16}$$

Second, for $F_{Y_0(1,0)|D=0,M=0}(y)$ we have

$$F^{-1}_{U_t|10}(F_{Y_0(1,0)|D=1,M=0}(y)) = h^{-1}(1,0,0;y), \tag{A.17}$$

using (A.1). Combining (A.16) and (A.17) yields,

$$h(1,0,1,h^{-1}(1,0,0;y)) = F^{-1}_{Y_1(1,0)|D=1,M=0} \circ F_{Y_0(1,0)|D=1,M=0}(y). \tag{A.18}$$

The left sides of (A.15) and (A.18) are equal. In contrast to (A.15), (A.18) contains only distributions that can be identified from observable data. In particular, $F_{Y_t(1,0)|D=1,M=0}(y) = \Pr(Y_t(1,0) \leq y|D=1,M=0) = \Pr(Y_t \leq y|D=1,M=0)$. Accordingly, we can identify $F^{-1}_{Y_1(1,0)|D=0,M=0} \circ F_{Y_0(1,0)|D=0,M=0}(y)$ by $Q_{10}(y) \equiv F^{-1}_{Y_1|D=1,M=0} \circ F_{Y_0|D=1,M=0}(y)$.

Parsing Y_0 through $Q_{10}(\cdot)$ in the non-treated group without mediator gives

$$\begin{aligned}
&E[Q_{10}(Y_0)|D=0,M=0] \\
&\quad = E[F^{-1}_{Y_1|D=1,M=0} \circ F_{Y_0|D=1,M=0}(Y_0)|D=0,M=0], \\
&\quad = E[F^{-1}_{Y_1(1,0)|D=1,M=0} \circ F_{Y_0(1,0)|D=1,M=0}(Y_0(0,0))|D=0,M=0], \\
&\quad \overset{A1,A3a}{=} E[h(1,0,1,h^{-1}(1,0,0;Y_0(0,0)))|D=0,M=0], \\
&\quad \overset{A2}{=} E[h(1,0,1,h^{-1}(1,0,0;Y_0(1,0)))|D=1,M=0], \\
&\quad \overset{A1,A3b}{=} E[F^{-1}_{Y_1(1,0)|D=0,M=0} \circ F_{Y_0(1,0)|D=0,M=0}(Y_0(1,0))|D=0,M=0], \\
&\quad = E[Y_1(1,0)|D=0,M=0] = E[Y_1(1,0)|D=0,M(0)=0],
\end{aligned} \tag{A.19}$$

which has data support because of Assumption 4b.

A.4 Quantile direct effect under d = 0 conditional on D = 0 and M(0) = 0

In the following, we prove that

$$\begin{aligned}\theta_1^{0,0}(q,0) &= F^{-1}_{Y_1(1,0)|D=0,M(0)=0}(q) - F^{-1}_{Y_1(0,0)|D=0,M(0)=0}(q),\\ &= F^{-1}_{Q_{10}(Y_0)|D=0,M=0}(q) - F^{-1}_{Y_1|D=0,M=0}(q).\end{aligned}$$

For this purpose, we have to show that

$$F_{Y_1(1,0)|D=0,M(0)=0}(y) = F_{Q_{10}(Y_0)|D=0,M=0}(y) \text{ and} \tag{A.20}$$

$$F_{Y_1(0,0)|D=0,M(0)=0}(y) = F_{Y_1|D=0,M=0}(y), \tag{A.21}$$

which is sufficient to show that the quantiles are also identified. We can show (A.21) using the observational rule $F_{Y_1(0,0)|D=0,M(0)=0}(y) = F_{Y_1|D=0,M=0}(y) = E[1\{Y_1 \leq y\}|D=0,M=0]$.

Furthermore, in analogy to (A.19), we obtain

$$\begin{aligned}&F_{Q_{10}(Y_0)|D=0,M=0}(y)\\ &\quad= E[1\{Q_{10}(Y_0) \leq y\}|D=0,M=0],\\ &\quad= E[1\{F^{-1}_{Y_1|D=1,M=0} \circ F_{Y_0|D=1,M=0}(Y_0) \leq y\}|D=0,M=0],\\ &\quad= E[1\{Y_1(1,0) \leq y\}|D=0,M=0],\\ &\quad= F_{Y_1(1,0)|D=0,M(0)=0}(y),\end{aligned}$$

which proves (A.20).

A.5 Average direct effect under d = 0 conditional on D = 0 and M(0) = 1

In the following, we show that $\theta_1^{0,1}(0) = E[Y_1(1,1) - Y_1(0,1)|D = 0, M(0) = 1] = E[Q_{11}(Y_0) - Y_1|D = 0, M = 1]$. Using the observational rule, we obtain $E[Y_1(0,1)|D = 0, M(0) = 1] = E[Y_1|D = 0, M = 1]$. Accordingly, we have to show that $E[Y_1(1,1)|D = 0, M(0) = 1] = E[Q_{11}(Y_0)|D = 0, M = 1]$ to finish the proof.

Under Assumptions 1 and 3c, the conditional potential outcome distribution function equals

$$\begin{aligned} F_{Y_t(d,0)|D=1,M=0}(y) &\overset{A1}{=} \Pr(h(d,m,t,U_t) \leq y|D = 0, M = 1, T = t), \\ &= \Pr(U_t \leq h^{-1}(d,m,t;y)|D = 0, M = 1, T = t), \\ &\overset{A3c}{=} \Pr(U_t \leq h^{-1}(d,m,t;y)|D = 0, M = 1), \\ &\overset{A3c}{=} \Pr(U_{t'} \leq h^{-1}(d,m,t;y)|D = 0, M = 1), \\ &= F_{U_{t'}|01}(h^{-1}(d,m,t;y)), \end{aligned} \tag{A.22}$$

for $d, t, t' \in \{0,1\}$. We use these quantities in the following.

First, evaluating $F_{Y_1(1,1)|D=0,M=1}(y)$ at $h(1,1,1,u)$ gives

$$F_{Y_1(1,1)|D=0,M=1}(h(1,1,1,u)) = F_{U_t|01}(h^{-1}(1,1,1;h(1,1,1,u))) = F_{U_t|01}(u),$$

for any $t \in \{0,1\}$. Applying $F^{-1}_{Y_1(1,1)|D=0,M=1}(q)$ to both sides, we have

$$h(1,1,1,u) = F^{-1}_{Y_1(1,1)|D=0,M=1}(F_{U_t|01}(u)). \tag{A.23}$$

Second, for $F_{Y_0(1,1)|D=0,M=1}(y)$ we have

$$F^{-1}_{U_t|01}(F_{Y_0(1,1)|D=0,M=1}(y)) = h^{-1}(1,1,0;y). \tag{A.24}$$

Combining (A.23) and (A.24) yields,

$$h(1,1,1,h^{-1}(1,1,0;y)) = F^{-1}_{Y_1(1,1)|D=0,M=1} \circ F_{Y_0(1,1)|D=0,M=1}(y). \tag{A.25}$$

Note that $h(1,1,1,h^{-1}(1,1,0;y))$ maps the period 1 (potential) outcome of an cross-sectional observation unit with the outcome y in period 0 under treatment with the mediator. Accordingly, $E[F^{-1}_{Y_1(1,1)|D=0,M=1} \circ F_{Y_0(1,1)|D=0,M=1}(Y_0)|D=0,M=1] = E[Y_1(1,1)|D=0,M=1]$. We can identify $F_{Y_0(1,1)|D=0,M=1}(y) = F_{Y_0|D=0,M=1}(y)$ under Assumption 2, but we cannot identify $F_{Y_1(1,1)|D=0,M=1}(y)$. We show in the following that we can identify the overall quantile-quantile transform $F^{-1}_{Y_1(1,1)|D=0,M=1} \circ F_{Y_0(1,1)|D=0,M=1}(y)$ under the additional Assumption 3d.

Under Assumptions 1 and 3d, the conditional potential outcome distribution function equals

$$\begin{aligned} F_{Y_t(d,1)|D=1,M=1}(y) &\overset{A1}{=} \Pr(h(d,m,t,U_t) \leq y|D=1,M=1,T=t), \\ &= \Pr(U_t \leq h^{-1}(d,m,t;y)|D=1,M=1,T=t), \\ &\overset{A3d}{=} \Pr(U_t \leq h^{-1}(d,m,t;y)|D=1,M=1), \\ &\overset{A3d}{=} \Pr(U_{t'} \leq h^{-1}(d,m,t;y)|D=1,M=1), \\ &= F_{U_{t'}|11}(h^{-1}(d,m,t;y)), \end{aligned} \tag{A.26}$$

for $d,t,t' \in \{0,1\}$. We repeat similar steps as above. First, evaluating $F_{Y_1(1,1)|D=1,M=1}(y)$ at $h(1,1,1,u)$ gives

$$F_{Y_1(1,1)|D=1,M=1}(h(1,1,1,u)) = F_{U_t|11}(h^{-1}(1,1,1;h(1,1,1,u))) = F_{U_t|11}(u),$$

for any $t \in \{0,1\}$. Applying $F^{-1}_{Y_1(1,1)|D=1,M=1}(q)$ to both sides, we have

$$h(1,1,1,u) = F^{-1}_{Y_1(1,1)|D=1,M=1}(F_{U_t|11}(u)). \tag{A.27}$$

Second, for $F_{Y_0(1,1)|D=1,M=1}(y)$ we have

$$F^{-1}_{U_t|11}(F_{Y_0(1,1)|D=1,M=1}(y)) = h^{-1}(1,1,1;y). \tag{A.28}$$

Combining (A.27) and (A.28) yields,

$$h(1,1,1,h^{-1}(1,1,0;y)) = F^{-1}_{Y_1(1,1)|D=1,M=1} \circ F_{Y_0(1,1)|D=1,M=1}(y). \tag{A.29}$$

The left sides of (A.25) and (A.29) are equal. In contrast to (A.25), (A.29) contains only distributions that can be identified from observable data. In particular, $F_{Y_t(1,1)|D=1,M=1}(y) = \Pr(Y_t(1,1) \leq y|D=1,M=1) = \Pr(Y_t \leq y|D=1,M=1)$. Accordingly, we can identify $F^{-1}_{Y_1(1,1)|D=0,M=1} \circ F_{Y_0(1,1)|D=0,M=1}(y)$ by $Q_{11}(y) \equiv F^{-1}_{Y_1|D=1,M=1} \circ F_{Y_0|D=1,M=1}(y)$.

Parsing Y_0 through $Q_{11}(\cdot)$ in the non-treated group with mediator gives

$$\begin{aligned}
&E[Q_{11}(Y_0)|D=0,M=1] \\
&\quad = E[F^{-1}_{Y_1|D=1,M=1} \circ F_{Y_0|D=1,M=1}(Y_0)|D=0,M=1], \\
&\quad = E[F^{-1}_{Y_1(1,1)|D=1,M=1} \circ F_{Y_0(1,1)|D=1,M=1}(Y_0(0,1))|D=0,M=1], \\
&\quad \overset{A1,A3d}{=} E[h(1,1,1,h^{-1}(1,1,0;Y_0(0,1)))|D=0,M=1], \\
&\quad \overset{A2}{=} E[h(1,1,1,h^{-1}(1,1,0;Y_0(0,0)))|D=0,M=1], \\
&\quad \overset{A1,A3c}{=} E[F^{-1}_{Y_1(1,1)|D=0,M=1} \circ F_{Y_0(1,1)|D=0,M=1}(Y_0(0,0))|D=0,M=1], \\
&\quad = E[Y_1(1,1)|D=0,M=1] = E[Y_1(1,1)|D=0,M(0)=1],
\end{aligned} \tag{A.30}$$

which has data support because of Assumption 4c.

A.6 Quantile direct effect under d = 0 conditional on D = 0 and M(0) = 1

In the following, we show that

$$\begin{aligned}
\theta_1^{0,1}(q,0) &= F^{-1}_{Y_1(1,1)|D=0,M(0)=1}(q) - F^{-1}_{Y_1(0,1)|D=0,M(0)=1}(q), \\
&= F^{-1}_{Q_{11}(Y_0)|D=0,M=1}(q) - F^{-1}_{Y_1|D=0,M=1}(q).
\end{aligned}$$

For this purpose, we have to prove that

$$F_{Y_1(1,1)|D=0,M(0)=1}(y) = F_{Q_{11}(Y_0)|D=0,M=1}(y) \text{ and} \tag{A.31}$$

$$F_{Y_1(0,1)|D=0,M(0)=1}(y) = F_{Y_1|D=0,M=1}(y), \tag{A.32}$$

which is sufficient to show that the quantiles are also identified. We can show (A.32) using the observational rule $F_{Y_1(0,1)|D=0,M(0)=1}(y) = F_{Y_1|D=0,M=1}(y) = E[1\{Y_1 \leq y\}|D=0, M=1]$.

In analogy to (A.30), we obtain

$$\begin{aligned}
&F_{Q_{11}(Y_0)|D=0,M=1}(y) \\
&\quad = E[1\{Q_{11}(Y_0) \leq y\}|D=0, M=1], \\
&\quad = E[1\{F^{-1}_{Y_1|D=1,M=1} \circ F_{Y_0|D=1,M=1}(Y_0) \leq y\}|D=0, M=1], \\
&\quad = E[1\{Y_1(1,1) \leq y\}|D=0, M=0], \\
&\quad = F_{Y_1(1,1)|D=0,M(0)=1}(y),
\end{aligned} \tag{A.33}$$

which proves (A.31).

A.7 Average direct effect under d = 1 conditional on D = 1 and M(1) = 1

In the following, we show that $\theta_1^{1,1}(1) = E[Y_1(1,1) - Y_1(0,1)|D=1, M(1)=1] = E[Y_1 - Q_{01}(Y_0)|D=1, M=1]$. Using the observational rule, we obtain $E[Y_1(1,1)|D=1, M(1)=1] = E[Y_1|D=1, M=1]$. Accordingly, we have to show that $E[Y_1(0,1)|D=1, M(1)=1] = E[Q_{01}(Y_0)|D=1, M=1]$ to finish the proof.

First, using (A.26) to evaluate $F_{Y_1(0,1)|D=1,M=1}(y)$ at $h(0,1,1,u)$ gives

$$F_{Y_1(0,1)|D=1,M=1}(h(0,1,1,u)) = F_{U_t|11}(h^{-1}(0,1,1;h(0,1,1,u))) = F_{U_t|11}(u),$$

for any $t \in \{0,1\}$. Applying $F^{-1}_{Y_1(0,1)|D=1,M=1}(q)$ to both sides, we have

$$h(0,1,1,u) = F^{-1}_{Y_1(0,1)|D=1,M=1}(F_{U_t|11}(u)). \tag{A.34}$$

Second, for $F_{Y_0(0,1)|D=0,M=1}(y)$ we obtain

$$F^{-1}_{U_t|11}(F_{Y_0(0,1)|D=1,M=1}(y)) = h^{-1}(0,1,0;y), \tag{A.35}$$

using (A.26). Combining (A.34) and (A.35) yields,

$$h(0,1,1,h^{-1}(0,1,0;y)) = F^{-1}_{Y_1(0,1)|D=1,M=1} \circ F_{Y_0(0,1)|D=1,M=1}(y). \tag{A.36}$$

Note that $h(0,1,1,h^{-1}(0,1,0;y))$ maps the period 1 (potential) outcome of an cross-sectional observation unit with the outcome y in period 0 under non-treatment with the mediator. Accordingly, $E[F^{-1}_{Y_1(1,1)|D=0,M=1} \circ F_{Y_0(1,1)|D=0,M=1}(Y_0)|D=0,M=1] = E[Y_1(1,1)|D=0,M=1]$. We can identify $F_{Y_0(1,1)|D=0,M=1}(y) = F_{Y_0|D=0,M=1}(y)$ under Assumption 2, but we cannot identify $F_{Y_1(1,1)|D=0,M=1}(y)$. We show in the following that we can identify the overall quantile-quantile transform $F^{-1}_{Y_1(1,1)|D=0,M=1} \circ F_{Y_0(1,1)|D=0,M=1}(y)$ under the additional Assumption 3c.

First, using (A.22) to evaluate $F_{Y_1(0,1)|D=0,M=1}(y)$ at $h(0,1,1,u)$ gives

$$F_{Y_1(0,1)|D=0,M=1}(h(0,1,1,u)) = F_{U_t|01}(h^{-1}(0,1,1;h(0,1,1,u))) = F_{U_t|01}(u),$$

for any $t \in \{0,1\}$. Applying $F^{-1}_{Y_1(0,1)|D=0,M=1}(q)$ to both sides, we have

$$h(0,1,1,u) = F^{-1}_{Y_1(0,1)|D=0,M=1}(F_{U_t|01}(u)). \tag{A.37}$$

Second, for $F_{Y_0(0,1)|D=0,M=1}(y)$ we obtain

$$F^{-1}_{U_t|01}(F_{Y_0(0,1)|D=0,M=1}(y)) = h^{-1}(0,1,1;y), \tag{A.38}$$

using (A.22). Combining (A.37) and (A.38) yields,

$$h(0,1,1,h^{-1}(0,1,0;y)) = F^{-1}_{Y_1(0,1)|D=0,M=1} \circ F_{Y_0(0,1)|D=0,M=1}(y). \tag{A.39}$$

The left sides of (A.36) and (A.39) are equal. In contrast to (A.36), (A.39) contains only distributions that can be identified from observable data. In particular, $F_{Y_t(0,1)|D=0,M=1}(y) = \Pr(Y_t(0,1) \leq y|D=0,M=1) = \Pr(Y_t \leq y|D=0,M=1)$. Accordingly, we can identify $F^{-1}_{Y_1(0,1)|D=1,M=1} \circ F_{Y_0(0,1)|D=1,M=1}(y)$ by $Q_{01}(y) \equiv F^{-1}_{Y_1|D=0,M=1} \circ F_{Y_0|D=0,M=1}(y)$.

Parsing Y_0 through $Q_{01}(\cdot)$ in the treated group with mediator gives

$$\begin{aligned}
&E[Q_{01}(Y_0)|D=1,M=1] \\
&\quad = E[F^{-1}_{Y_1|D=0,M=1} \circ F_{Y_0|D=0,M=1}(Y_0)|D=1,M=1], \\
&\quad = E[F^{-1}_{Y_1(0,1)|D=0,M=1} \circ F_{Y_0(0,1)|D=0,M=1}(Y_0(1,1))|D=1,M=1], \\
&\quad \overset{A1,A3c}{=} E[h(0,1,1,h^{-1}(0,1,0;Y_0(1,1)))|D=1,M=1], \\
&\quad \overset{A2}{=} E[h(0,1,1,h^{-1}(0,1,0;Y_0(0,1)))|D=1,M=1], \\
&\quad \overset{A1,A3d}{=} E[F^{-1}_{Y_1(0,1)|D=1,M=1} \circ F_{Y_0(0,1)|D=1,M=1}(Y_0(0,1))|D=1,M=1], \\
&\quad = E[Y_1(0,1)|D=1,M=1] = E[Y_1(0,1)|D=1,M(1)=1],
\end{aligned} \tag{A.40}$$

which has data support under Assumption 4d.

A.8 Quantile direct effect under d = 1 conditional on D = 1 and M(1) = 1

In the following, we show that

$$\begin{aligned}
\theta^{1,1}_1(q,1) &= F^{-1}_{Y_1(1,1)|D=1,M(1)=1}(q) - F^{-1}_{Y_1(0,1)|D=1,M(1)=1}(q), \\
&= F^{-1}_{Y_1|D=1,M=1}(q) - F^{-1}_{Q_{01}(Y_0)|D=1,M=1}(q).
\end{aligned}$$

For this purpose, we have to prove that

$$F_{Y_1(1,1)|D=1,M(1)=1}(y) = F_{Y_1|D=1,M=1}(y) \text{ and} \tag{A.41}$$

$$F_{Y_1(0,1)|D=1,M(1)=1}(y) = F_{Q_{01}(Y_0)|D=1,M=1}(y), \tag{A.42}$$

which is sufficient to show that the quantiles are also identified. We can show (A.41) using the observational rule $F_{Y_1(1,1)|D=1,M(1)=1}(y) = F_{Y_1|D=1,M=1}(y) = E[1\{Y_1 \leq y\}|D=1, M=1]$.

In analogy to (A.40), we obtain

$$\begin{aligned}
&F_{Q_{01}(Y_0)|D=1,M=1}(y) \\
&\quad = E[1\{Q_{01}(Y_0) \leq y\}|D=1, M=1], \\
&\quad = E[1\{F^{-1}_{Y_1|D=0,M=1} \circ F_{Y_0|D=0,M=1}(Y_0) \leq y\}|D=1, M=1], \\
&\quad = E[1\{Y_1(0,1) \leq y\}|D=1, M=0], \\
&\quad = F_{Y_1(0,1)|D=1,M(1)=1}(y),
\end{aligned}$$

which proves (A.42).

B Proof of Equations (1) and (2)

The average total effect for the entire population is identified by,

$$\begin{aligned}
\Delta_1 &= E[Y_1(1, M(1))] - E[Y_1(0, M(0))], \\
&\overset{A5}{=} E[Y_1(1, M(1))|D=1] - E[Y_1(0, M(0))|D=0], \\
&= E[Y_1|D=1] - E[Y_1|D=0],
\end{aligned}$$

where the first equality is the definition of Δ_1, the second equality hold by Assumption 5, and the last equality holds by the observational rule.

We define the conditional distribution $F_{Y_1|D=d}(y) = \Pr(Y_1 \leq y|D=d)$ and $F^{-1}_{Y_1|D=d}(q) = \inf\{y : F_{Y_1|D=d}(y) \geq q\}$. We can show the identification of the total QTE for the entire population $\Delta_1(q) = F^{-1}_{Y_1|D=1}(q) - F^{-1}_{Y_1|D=0}(q)$ when we show that

$F_{Y_1(1,M(1))}(y) = F_{Y_1|D=1}(y)$ and $F_{Y_1(0,M(0))}(y) = F_{Y_1|D=0}(y)$. Using Assumption 5 and the observational rule gives,

$$\begin{aligned} F_{Y_1(1,M(1))}(y) &= \Pr(Y_1(1, M(1)) \leq y), \\ &\overset{A5}{=} \Pr(Y_1(1, M(1)) \leq y | D = 1), \\ &= \Pr(Y_1 \leq y | D = 1) = F_{Y_1|D=1}(y), \end{aligned}$$

and

$$\begin{aligned} F_{Y_1(0,M(0))}(y) &= \Pr(Y_1(0, M(0)) \leq y), \\ &\overset{A5}{=} \Pr(Y_1(0, M(0)) \leq y | D = 0), \\ &= \Pr(Y_1 \leq y | D = 0) = F_{Y_1|D=0}(y), \end{aligned}$$

which finishes the proof.

By Assumption 5, the share of a type τ conditional on D corresponds to p_τ (in the population), as D is randomly assigned. This implies that $p_{1|1} = p_{n1} + p_{ap}$, $p_{1|0} = p_{n1} + p_{an}$, $p_{0|1} = p_{n0} + p_{an}$, and $p_{0|0} = p_{n0} + p_{ap}$. Under Assumption 6, $p_{an} = 0$, which finishes the proof of equation (1).

Furthermore, $E[Y_t(d, m)|\tau, D = 1] = E[Y_t(d, m)|\tau, D = 0] = E[Y_t(d, m)|\tau]$ due to the independence of D and the potential outcomes as well as the types τ (which are a deterministic function of $M(d)$) under Assumption 5. It follows that conditioning on D is not required on the right hand side of the following equation, which expresses the mean outcome conditional $D = 0$ and $M = 0$ as weighted average of the mean potential outcomes of affected positively and not-affected at 0:

$$\begin{aligned} &E[Y_t|D = 0, M = 0] \\ &\quad = \frac{p_{n0}}{p_{n0} + p_{ap}} E[Y_t(0, 0)|\tau = n0] + \frac{p_{ap}}{p_{n0} + p_{ap}} E[Y_t(0, 0)|\tau = ap]. \end{aligned} \tag{B.1}$$

Only affected positively and not-affected at 0 satisfy $M(0) = 0$ and thus make up

the group with $D = 0$ and $M = 0$. After some rearrangements we obtain

$$\begin{aligned} &E[Y_t(0,0)|\tau = n0] - E[Y_t(0,0)|\tau = ap] \\ &\quad = \frac{p_{n0} + p_{ap}}{p_{ap}} \left\{ E[Y_t(0,0)|\tau = n0] - E[Y_t|D = 0, M = 0] \right\}. \end{aligned} \tag{B.2}$$

Next, we consider observations with $D = 1$ and $M = 0$, which might consist of both not-affected at 0 and affected negatively, as $M(1) = 0$ for both types. However, by Assumption 6, affected negatively are ruled out, such that the mean outcome given $D_1 = 1$ and $M_1 = 0$ is determined by not-affected at 0 only:

$$E[Y_t|D = 1, M = 0] \overset{A5,A6}{=} E[Y_t(1,0)|\tau = n0]. \tag{B.3}$$

Furthermore, by Assumption 2,

$$E[Y_0(0,0)|\tau = n0] \overset{A2}{=} E[Y_0(1,0)|\tau = n0] \overset{A5,A6}{=} E[Y_0|D = 1, M = 0].$$

Similarly to (B.1) for the not-affected at 0 and affected positively, consider the mean outcome given $D = 1$ and $M = 1$, which is made up by not-affected at 1 and affected positively (the types with $M(1) = 1$)

$$\begin{aligned} &E[Y_t|D = 1, M = 1] \\ &\quad = \frac{p_{n1}}{p_{n1} + p_{ap}} E[Y_t(1,1)|\tau = n1] + \frac{p_{ap}}{p_{n1} + p_{ap}} E[Y_t(1,1)|\tau = ap]. \end{aligned} \tag{B.4}$$

After some rearrangements we obtain

$$\begin{aligned} &E[Y_t(1,1)|\tau = n1] - E[Y_t(1,1)|\tau = ap] \\ &\quad = \frac{p_{n1} + p_{ap}}{p_{ap}} \left\{ E[Y_t(1,1)|\tau = n1] - E[Y_t|D = 1, M = 1] \right\}. \end{aligned} \tag{B.5}$$

By Assumptions 5 and 6,

$$E[Y_t|D = 0, M = 1] = E[Y_t(0,1)|\tau = n1]. \tag{B.6}$$

Now consider (B.5) for period $T = 0$, and note that by Assumption 2, $E[Y_0(1,1)|\tau = n1] = E[Y_0(0,0)|\tau = n1] = E[Y_0(0,1)|\tau = n1]$ and $E[Y_0(1,1)|\tau = ap] = E[Y_0(0,0)|\tau = ap]$.

Combining (B.4), (B.6), and the law of iterative expectations (LIE) gives

$$
\begin{aligned}
&E[Y_0|D=1] \\
&\quad \overset{LIE}{=} E[Y_0|D=1, M=1] \cdot p_{1|1} + E[Y_0|D=1, M=0] \cdot p_{0|1}, \\
&\quad = E[Y_0(1,1)|\tau = ap] \cdot p_{ap} + E[Y_0(1,1)|\tau = n1] \cdot p_{n1} + E[Y_0(1,0)|\tau = n0] \cdot p_{n0}, \\
&\quad \overset{A2}{=} E[Y_0(1,1)|\tau = ap] \cdot p_{ap} + E[Y_0(1,1)|\tau = n1] \cdot p_{n1} + E[Y_0(0,0)|\tau = n0] \cdot p_{n0}.
\end{aligned}
$$

Likewise, combining (B.1) and (B.3) gives

$$
\begin{aligned}
&E[Y_0|D=0] \\
&\quad \overset{LIE}{=} E[Y_0|D=0, M=1] \cdot p_{1|0} + E[Y_0|D_0=1, M=0] \cdot p_{0|0}, \\
&\quad = E[Y_0(0,1)|\tau = n1] \cdot p_{n1} + E[Y_0(0,0)|\tau = ap] \cdot p_{ap} + E[Y_0(0,0)|\tau = n0] \cdot p_{n0}, \\
&\quad \overset{A2}{=} E[Y_0(1,1)|\tau = n1] \cdot p_{n1} + E[Y_0(0,0)|\tau = ap] \cdot p_{ap} + E[Y_0(0,0)|\tau = n0] \cdot p_{n0}.
\end{aligned}
$$

Accordingly,

$$
\frac{E[Y_0|D=1] - E[Y_0|D=0]}{p_{1|1} - p_{1|0}} = E[Y_0(1,1)|\tau = ap] - E[Y_0(0,0)|\tau = ap] \overset{A2}{=} 0,
$$

which proves equation (2). Accordingly, $E[Y_0|D=1] - E[Y_0|D=0] = 0$ is a testable implication of Assumption 2, 5, and 6.

C Proof of Theorem 2

C.1 Average direct effect on the not-affected at 0

In the following, we show that $\theta_1^{n0} = E[Y_1(1,0) - Y_1(0,0)|\tau = n0] = E[Y_1 - Q_{00}(Y_0)|D=1, M=0]$. From (B.3), we obtain the first ingredient $E[Y_1(1,0)|\tau =$

$n0] = E[Y_1|D = 1, M = 0]$. Furthermore, from (A.9) we have $E[Q_{00}(Y_0)|D = 1, M = 0] = E[Y_1(0,0)|D = 1, M(1) = 0]$. Under Assumption 5 and 6,

$$E[Y_1(0,0)|D = 1, M(1) = 0] = E[Y_1(0,0)|D = 1, \tau = n0] = E[Y_1(0,0)|\tau = n0]. \tag{C.1}$$

C.2 Quantile direct effect on the not-affected at 0

We prove that

$$\begin{aligned}\theta_1^{n0}(q) &= F^{-1}_{Y_1(1,0)|n0}(q) - F^{-1}_{Y_1(0,0)|n0}(q), \\ &= F^{-1}_{Y_1|D=1,M=0}(q) - F^{-1}_{Q_{00}(Y_0)|D=1,M=0}(q).\end{aligned}$$

This requires showing that

$$F_{Y_1(1,0)|n0}(y) = F_{Y_1|D=1,M=0}(y) \text{ and} \tag{C.2}$$

$$F_{Y_1(0,0)|n0}(y) = F_{Q_{00}(Y_0)|D=1,M=0}(y). \tag{C.3}$$

Under Assumptions 5 and 6,

$$\begin{aligned}F_{Y_t|D=1,M=0}(y) &= E[1\{Y_t \leq y\}|D = 1, M = 0] \\ &\overset{A5,A6}{=} E[1\{Y_t(1,0) \leq y\}|\tau = n0] \\ &= F_{Y_t(1,0)|n0}(y),\end{aligned} \tag{C.4}$$

which proves (C.2). From (A.12), we have

$$F_{Q_{00}(Y_0)|D=1,M=0}(y) = F_{Y_1(0,0)|D=1,M(1)=0}(y) = E[1\{Y_1(0,0) \leq y\}|D = 1, M(1) = 0].$$

Under Assumption 5 and 6,

$$\begin{aligned}E[1\{Y_1(0,0) \leq y\}|D = 1, M(1) = 0] &\overset{A5,A6}{=} E[1\{Y_1(0,0) \leq y\}|\tau = n0] \\ &= F_{Y_1(0,0)|n0}(y),\end{aligned} \tag{C.5}$$

which proves (C.3).

C.3 Average direct effect under d = 0 on affected positively

In the following, we show that

$$\begin{aligned}\theta_1^{ap}(0) =& E[Y_1(1,0) - Y_1(0,0)|\tau = ap],\\ =& \frac{p_{0|0}}{p_{0|0} - p_{0|1}} E[Q_{10}(Y_0) - Y_1|D = 0, M = 0]\\ &- \frac{p_{0|1}}{p_{0|0} - p_{0|1}} E[Y_1 - Q_{00}(Y_0)|D = 1, M = 0].\end{aligned}$$

Plugging (C.1) in (B.1) under $T = 1$, we obtain

$$\begin{aligned}E[Y_1|D = 0, M = 0] = & \frac{p_{n0}}{p_{n0} + p_{ap}} E[Q_{00}(Y_0)|D = 1, M = 0]\\ &+ \frac{p_{ap}}{p_{n0} + p_{ap}} E[Y_1(0,0)|\tau = ap].\end{aligned}$$

This allows identifying

$$\begin{aligned}E[Y_1(0,0)|\tau = ap] = & \frac{p_{0|0}}{p_{0|0} - p_{0|1}} E[Y_1|D = 0, M = 0]\\ &- \frac{p_{0|1}}{p_{0|0} - p_{0|1}} E[Q_{00}(Y_0)|D = 1, M = 0].\end{aligned} \tag{C.6}$$

Accordingly, we have to show the identification of $E[Y_1(1,0)|ap]$ to finish the proof. From (A.19) we have $E[Y_1(1,0)|D = 0, M = 0] = E[Q_{10}(Y_0)|D = 0, M = 0]$. Applying the law of iterative expectations, gives

$$\begin{aligned}E[Y_1(1,0)|D = 0, M = 0] =& \frac{p_{n0}}{p_{n0} + p_{ap}} E[Y_1(1,0)|D = 0, M = 0, \tau = n0]\\ &+ \frac{p_{ap}}{p_{n0} + p_{ap}} E[Y_1(1,0)|D = 0, M = 0, \tau = ap],\\ \overset{A5}{=}& \frac{p_{n0}}{p_{n0} + p_{ap}} E[Y_1(1,0)|\tau = n0] + \frac{p_{ap}}{p_{n0} + p_{ap}} E[Y_1(1,0)|\tau = ap].\end{aligned}$$

After some rearrangements and using (B.3), we obtain

$$E[Y_1(1,0)|\tau = ap] = \frac{p_{n0}+p_{ap}}{p_{ap}} E[Q_{10}(Y_0)|D=0, M=0] - \frac{p_{n0}}{p_{ap}} E[Y_1|D=1, M=0].$$

This gives

$$\begin{aligned} E[Y_1(1,0)|\tau = ap] = \quad & \frac{p_{0|0}}{p_{0|0}-p_{0|1}} E[Q_{10}(Y_0)|D=0, M=0] \\ & - \frac{p_{0|1}}{p_{0|0}-p_{0|1}} E[Y_1|D=1, M=0], \end{aligned} \tag{C.7}$$

using $p_{n0} = p_{0|1}$, and $p_{ap} + p_{n0} = p_{0|0}$.

C.4 Quantile direct effect under d = 0 on affected positively

We show that

$$\begin{aligned} F_{Y_1(1,0)|ap}(y) &= \frac{p_{0|0}}{p_{0|0}-p_{0|1}} F_{Q_{10}(Y_0)|D=0,M=0}(y) - \frac{p_{0|1}}{p_{0|0}-p_{0|1}} F_{Y_1|D=1,M=0}(y) \text{ and} \\ F_{Y_1(0,0)|ap}(y) &= \frac{p_{0|0}}{p_{0|0}-p_{0|1}} F_{Y_1|D=0,M=0}(y) - \frac{p_{0|1}}{p_{0|0}-p_{0|1}} F_{Q_{00}(Y_0)|D=1,M=0}(y), \end{aligned}$$

which proves that $\theta_1^{ap}(q,0) = F^{-1}_{Y_1(1,0)|ap}(q) - F^{-1}_{Y_1(0,0)|ap}(q)$ is identified.

From (A.20), we have $F_{Y_1(1,0)|D=0,M(0)=0}(y) = F_{Q_{10}(Y_0)|D=0,M=0}(y)$. Applying the law of iterative expectations gives

$$\begin{aligned} F_{Y_1(1,0)|D=0,M(0)=0}(y) =& \frac{p_{n0}}{p_{n0}+p_{ap}} F_{Y_1(1,0)|D=0,M(0)=0,\tau=n0}(y) \\ & + \frac{p_{ap}}{p_{n0}+p_{ap}} F_{Y_1(1,0)|D=0,M(0)=0,\tau=ap}(y), \\ \overset{A5}{=} & \frac{p_{n0}}{p_{n0}+p_{ap}} F_{Y_1(1,0)|n0}(y) + \frac{p_{ap}}{p_{n0}+p_{ap}} F_{Y_1(1,0)|ap}(y). \end{aligned}$$

Using (C.2) and rearranging the equation gives,

$$F_{Y_1(1,0)|ap}(y) = \frac{p_{0|0}}{p_{0|0}-p_{0|1}} F_{Q_{10}(Y_0)|D=0,M=0}(y) - \frac{p_{0|1}}{p_{0|0}-p_{0|1}} F_{Y_1|D=1,M=0}(y). \tag{C.8}$$

In analogy to (B.1), the outcome distribution under $D = 0$ and $M = 0$ equals

$$F_{Y_1|D=0,M=0}(y) = \frac{p_{n0}}{p_{n0} + p_{ap}} F_{Y_1(0,0)|n0}(y) + \frac{p_{ap}}{p_{n0} + p_{ap}} F_{Y_1(0,0)|ap}(y).$$

Using (C.3) and rearranging the equation gives

$$F_{Y_1(0,0)|ap}(y) = \frac{p_{0|0}}{p_{0|0} - p_{0|1}} F_{Y_1|D=0,M=0}(y) - \frac{p_{0|1}}{p_{0|0} - p_{0|1}} F_{Q_{00}(Y_0)|D=1,M=0}(y). \tag{C.9}$$

C.5 Average direct effect on the not-affected at 1

In the following, we show that $\theta_1^{n1} = E[Y_1(1,1) - Y_1(0,1)|\tau = n1] = E[Q_{11}(Y_0) - Y_1|D = 0, M = 1]$. From (B.6), we obtain the first ingredient $E[Y_1(0,1)|n1] = E[Y_1|D = 0, M = 1]$. Furthermore, from (A.30) we have $E[Q_{11}(Y_0)|D = 0, M = 1] = E[Y_1(1,1)|D = 0, M(0) = 1]$. Under Assumption 5 and 6,

$$E[Y_1(1,1)|D = 0, M(0) = 1] = E[Y_1(1,1)|D = 0, \tau = n1] = E[Y_1(1,1)|\tau = n1]. \tag{C.10}$$

C.6 Quantile direct effect on the not-affected at 1

We prove that

$$\begin{aligned}\theta_1^{n1}(q) &= F^{-1}_{Y_1(1,1)|n1}(q) - F^{-1}_{Y_1(0,1)|n1}(q), \\ &= F^{-1}_{Q_{11}(Y_0)|D=0,M=1}(q) - F^{-1}_{Y_1|D=0,M=1}(q).\end{aligned}$$

This requires showing that

$$F_{Y_1(1,1)|n1}(y) = F_{Q_{11}(Y_0)|D=0,M=1}(y) \text{ and} \tag{C.11}$$

$$F_{Y_1(0,1)|n1}(y) = F_{Y_1|D=0,M=1}(y). \tag{C.12}$$

Under Assumptions 5 and 6,

$$\begin{aligned} F_{Y_t|D=0,M=1}(y) &= E[1\{Y_t \leq y\}|D=0, M=1] \\ &\overset{A5,A6}{=} E[1\{Y_t(0,1) \leq y\}|\tau = n1] \\ &= F_{Y_t(0,1)|n1},(y). \end{aligned} \tag{C.13}$$

which proves (C.12). From (A.33), we have

$$F_{Q_{11}(Y_0)|D=0,M=1}(y) = F_{Y_1(1,1)|D=0,M(0)=1}(y) = E[1\{Y_1(1,1) \leq y\}|D=0, M(0)=1].$$

Under Assumption 5 and 6,

$$\begin{aligned} E[1\{Y_1(1,1) \leq y\}|D=0, M(0)=1] &\overset{A5,A6}{=} E[1\{Y_1(1,1) \leq y\}|\tau = n1] \\ &= F_{Y_1(1,1)|n1}(y), \end{aligned} \tag{C.14}$$

which proves (C.11).

C.7 Average direct effect under d = 1 on affected positively

In the following, we show that

$$\begin{aligned} \theta_1^{ap}(1) =& E[Y_1(1,1) - Y_1(0,1)|\tau = ap], \\ =& \frac{p_{1|1}}{p_{1|1} - p_{1|0}} E[Y_1 - Q_{01}(Y_0)|D=1, M=1] \\ & - \frac{p_{1|0}}{p_{1|1} - p_{1|0}} E[Q_{11}(Y_0) - Y_1|D=0, M=1]. \end{aligned}$$

Plugging (C.10) in (B.4), we obtain

$$\begin{aligned} E[Y_1|D=1, M=1] = & \frac{p_{n1}}{p_{n1} + p_{ap}} E[Q_{11}(Y_0)|D=0, M=1] \\ & + \frac{p_{ap}}{p_{n1} + p_{ap}} E[Y_1(1,1)|\tau = ap]. \end{aligned}$$

This allows identifying

$$\begin{aligned} E[Y_1(1,1)|\tau = ap] = \quad & \frac{p_{1|1}}{p_{1|1}-p_{1|0}} E[Y_1|D=1,M=1] \\ & - \frac{p_{1|0}}{p_{1|1}-p_{1|0}} E[Q_{11}(Y_0)|D=0,M=1]. \end{aligned} \tag{C.15}$$

From (A.40) we have $E[Y_1(0,1)|D=1,M=1] = E[Q_{01}(Y_0)|D=1,M=1]$. Applying the law of iterative expectations, gives

$$\begin{aligned} E[Y_1(0,1)|D=1,M=1] =& \frac{p_{n1}}{p_{n1}+p_{ap}} E[Y_1(0,1)|D=1,M=1,\tau=n1] \\ & + \frac{p_{ap}}{p_{n1}+p_{ap}} E[Y_1(0,1)|D=1,M=1,\tau=ap], \\ \overset{A5}{=} & \frac{p_{n1}}{p_{n1}+p_{ap}} E[Y_1(0,1)|\tau=n1] + \frac{p_{ap}}{p_{n1}+p_{ap}} E[Y_1(0,1)|\tau=ap]. \end{aligned}$$

After some rearrangements and using (B.6), we obtain

$$E[Y_1(0,1)|\tau=ap] = \frac{p_{n1}+p_{ap}}{p_{ap}} E[Q_{01}(Y_0)|D=1,M=1] - \frac{p_{n1}}{p_{ap}} E[Y_1|D=0,M=1].$$

This gives

$$\begin{aligned} E[Y_1(0,1)|\tau = ap] = \quad & \frac{p_{1|1}}{p_{1|1}-p_{1|0}} E[Q_{01}(Y_0)|D=1,M=1] \\ & - \frac{p_{1|0}}{p_{1|1}-p_{1|0}} E[Y_1|D=0,M=1], \end{aligned} \tag{C.16}$$

with $p_{n1} = p_{1|0}$, and $p_{ap} + p_{n1} = p_{1|1}$.

C.8 Quantile direct effect under d = 1 on affected positively

We show that

$$\begin{aligned} F_{Y_1(1,1)|ap}(y) &= \frac{p_{1|1}}{p_{1|1}-p_{1|0}} F_{Y_1|D=1,M=1}(y) - \frac{p_{1|0}}{p_{1|1}-p_{1|0}} F_{Q_{11}(Y_0)|D=0,M=1}(y) \text{ and} \\ F_{Y_1(0,1)|ap}(y) &= \frac{p_{1|1}}{p_{1|1}-p_{1|0}} F_{Q_{01}(Y_0)|D=1,M=1}(y) - \frac{p_{1|0}}{p_{1|1}-p_{1|0}} F_{Y_1|D=0,M=1}(y), \end{aligned}$$

which proves that $\theta_1^{ap}(q,1) = F^{-1}_{Y_1(1,1)|ap}(q) - F^{-1}_{Y_1(0,1)|ap}(q)$ is identified.

In analogy to (B.4), the outcome distribution under $D = 0$ and $M = 0$ equals:

$$F_{Y_1|D=1,M=1}(y) = \frac{p_{n1}}{p_{n1} + p_{ap}} F_{Y_1(1,1)|n1}(y) + \frac{p_{ap}}{p_{n1} + p_{ap}} F_{Y_1(1,1)|ap}(y).$$

Using (C.11) and rearranging the equation gives

$$F_{Y_1(1,1)|ap}(y) = \frac{p_{1|1}}{p_{1|1} - p_{1|0}} F_{Y_1|D=1,M=1}(y) - \frac{p_{1|0}}{p_{1|1} - p_{1|0}} F_{Q_{11}(Y_0)|D=0,M=1}(y). \quad \text{(C.17)}$$

From (A.42), we have $F_{Y_1(0,1)|D=1,M(1)=1}(y) = F_{Q_{01}(Y_0)|D=1,M=1}(y)$. Applying the law of iterative expectations gives

$$\begin{aligned} F_{Y_1(0,1)|D=1,M(1)=1}(y) =& \frac{p_{n1}}{p_{n1} + p_{ap}} F_{Y_1(0,1)|D=1,M(1)=1,\tau=n1}(y) \\ &+ \frac{p_{ap}}{p_{n1} + p_{ap}} F_{Y_1(0,1)|D=1,M(1)=1,\tau=ap}(y), \\ \overset{A5}{=}& \frac{p_{n1}}{p_{n1} + p_{ap}} F_{Y_1(0,1)|n1}(y) + \frac{p_{ap}}{p_{n1} + p_{ap}} F_{Y_1(0,1)|ap}(y). \end{aligned}$$

Using (C.12) and rearranging the equation gives,

$$F_{Y_1(0,1)|ap}(y) = \frac{p_{1|1}}{p_{1|1} - p_{1|0}} F_{Q_{01}(Y_0)|D=1,M=1}(y) - \frac{p_{1|0}}{p_{1|1} - p_{1|0}} F_{Y_1|D=0,M=1}(y). \quad \text{(C.18)}$$

D Proof of Theorem 3

D.1 Average treatment effect on the affected positively

In (C.15) and (C.6), we show that

$$\begin{aligned} \theta_1^{ap} =& E[Y_1(1,1) - Y_1(0,0)|\tau = ap], \\ =& \frac{p_{1|1}}{p_{1|1} - p_{1|0}} E[Y_1|D = 1, M = 1] - \frac{p_{1|0}}{p_{1|1} - p_{1|0}} E[Q_{11}(Y_0)|D = 0, M = 1] \\ &- \frac{p_{0|0}}{p_{0|0} - p_{0|1}} E[Y_1|D = 0, M = 0] + \frac{p_{0|1}}{p_{0|0} - p_{0|1}} E[Q_{00}(Y_0)|D = 1, M = 0]. \end{aligned}$$

D.2 Quantile treatment effect on the affected positively

In (C.17) and (C.9), we show that $F_{Y_1(1,1)|ap}(y)$ and $F_{Y_1(0,0)|ap}(y)$ are identified. Accordingly, $\Delta_1^{ap}(q) = F^{-1}_{Y_1(1,1)|ap}(q) - F^{-1}_{Y_1(0,0)|ap}(q)$ is identified.

D.3 Average indirect effect under $d = 0$ on affected positively

In (C.16) and (C.6), we show that

$$\begin{aligned}\delta_1^{ap}(0) =& E[Y_1(0,1) - Y_1(0,0)|\tau = ap],\\ =& \frac{p_{1|1}}{p_{1|1} - p_{1|0}} E[Q_{01}(Y_0)|D = 1, M = 1] - \frac{p_{1|0}}{p_{1|1} - p_{1|0}} E[Y_1|D = 0, M = 1]\\ & - \frac{p_{0|0}}{p_{0|0} - p_{0|1}} E[Y_1|D = 0, M = 0] + \frac{p_{0|1}}{p_{0|0} - p_{0|1}} E[Q_{00}(Y_0)|D = 1, M = 0].\end{aligned}$$

D.4 Quantile indirect effect under $d = 0$ on affected positively

In (C.18) and (C.9), we show that $F_{Y_1(0,1)|ap}(y)$ and $F_{Y_1(0,0)|ap}(y)$ are identified. Accordingly, $\delta_1^{ap}(q, 0) = F^{-1}_{Y_1(0,1)|ap}(q) - F^{-1}_{Y_1(0,0)|ap}(q)$ is identified.

D.5 Average indirect effect under $d = 1$ on affected positively

In (C.15) and (C.7), we show that

$$\begin{aligned}\delta_1^{ap}(1) =& E[Y_1(1,1) - Y_1(1,0)|\tau = ap],\\ =& \frac{p_{1|1}}{p_{1|1} - p_{1|0}} E[Y_1|D = 1, M = 1] - \frac{p_{1|0}}{p_{1|1} - p_{1|0}} E[Q_{11}(Y_0)|D = 0, M = 1]\\ & - \frac{p_{0|0}}{p_{0|0} - p_{0|1}} E[Q_{10}(Y_0)|D = 0, M = 0] + \frac{p_{0|1}}{p_{0|0} - p_{0|1}} E[Y_1|D = 1, M = 0].\end{aligned}$$

D.6 Quantile indirect effect under $d = 1$ on affected positively

In (C.17) and (C.8), we show that $F_{Y_1(1,1)|ap}(y)$ and $F_{Y_1(1,0)|ap}(y)$ are identified. Accordingly, $\delta_1^{ap}(q,1) = F^{-1}_{Y_1(1,1)|ap}(q) - F^{-1}_{Y_1(1,0)|ap}(q)$ is identified.

E Simulation study

To shape the intuition for our identification results, this appendix presents a brief simulation based on the following data generating process (DGP):

$$T \sim Binom(0.5),\ D \sim Binom(0.5),\ U \sim Unif(-1,1),\ V \sim N(0,1)$$

independent of each other, and

$$M = I\{D + U + V > 0\}, \qquad Y_T = \Lambda((1 + D + M + D \cdot M) \cdot T + U).$$

Treatment D as well as the observed time period T are randomized and binomially distributed with a 50% chance of being 1 or 0, while the mediator-outcome association is confounded due to the unobserved time constant heterogeneity U (implying $U_0 = U_1$). The potential outcome in period 1 is given by $Y_1(d, M(d')) = \Lambda((1 + d + M(d') + d \cdot M(d')) + U)$, where Λ denotes a link function. If the latter corresponds to the identity function, our model is linear and implies a homogeneous time trend T equal to 1. If Λ is nonlinear, the time trend is heterogeneous, which invalidates the common trend assumption of DiD models. M is not only a function of D and U, but also of the unobserved random term V, which guarantees common support w.r.t. U, see Assumption 4. Affected positively, not-affected at 1, and not-affected at 0 satisfy, respectively: $ap = I\{U + V \leq 0, 1 + U + V > 0\}$, $n1 = I\{U + V > 0\}$, and $n0 = I\{1 + U + V \leq 0\}$.

In the simulations with 1,000 replications, we consider two sample sizes ($N = 1,000, 4,000$) and investigate the behaviour of our CiC approach as well as the

DiD approach of Deuchert, Huber, and Schelker (2019) in both a linear (Λ equal to identity function) and nonlinear outcome model where Λ equals the exponential function. The latter implies that a specific ceteris paribus change in a right hand variable entails a specific percentage change in the outcome (rather than a specific level change as in the linear model). To implement the CiC estimators in the simulations as well as the application in Section 4, we make use of the 'cic' command in the `qte` R-package by Callaway (2016) with its default values.

Table E.1: Linear model with random treatment

	$\hat{\theta}_1^{n0}$	$\hat{\theta}_1^{n1}$	$\hat{\Delta}_1^{ap}$	$\hat{\theta}_1^{ap}(1)$	$\hat{\theta}_1^{ap}(0)$	$\hat{\delta}_1^{ap}(1)$	$\hat{\delta}_1^{ap}(0)$
A. Changes-in-Changes							
				N=1,000			
bias	0.00	-0.00	-0.01	-0.01	-0.01	-0.00	-0.01
sd	0.11	0.08	0.23	0.10	0.13	0.27	0.27
rmse	0.11	0.08	0.23	0.10	0.13	0.27	0.27
true	1.00	2.00	3.00	2.00	1.00	2.00	1.00
relr	0.11	0.04	0.08	0.05	0.13	0.14	0.27
				N=4,000			
bias	-0.00	-0.00	0.00	-0.00	-0.01	0.01	0.01
sd	0.06	0.04	0.12	0.05	0.07	0.14	0.14
rmse	0.06	0.04	0.12	0.05	0.07	0.14	0.14
true	1.00	2.00	3.00	2.00	1.00	2.00	1.00
relr	0.06	0.02	0.04	0.02	0.07	0.07	0.14
B. Difference-in-Differences							
				N=1,000			
bias	0.01	-0.00	-0.01	-0.01	0.00	-0.02	0.00
sd	0.11	0.09	0.14	0.14	0.12	0.19	0.10
rmse	0.11	0.09	0.14	0.14	0.12	0.19	0.10
true	1.00	2.00	3.00	2.00	1.00	2.00	1.00
relr	0.11	0.04	0.05	0.07	0.12	0.10	0.10
				N=4,000			
bias	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00
sd	0.06	0.04	0.07	0.07	0.06	0.10	0.05
rmse	0.06	0.04	0.07	0.07	0.06	0.10	0.05
true	1.00	2.00	3.00	2.00	1.00	2.00	1.00
relr	0.06	0.02	0.02	0.04	0.06	0.05	0.05

Note: 'bias', 'sd', and 'rmse' provide the bias, standard deviation, and root mean squared error of the respective estimator. 'true' and 'relr' are the respective true effect as well as the root mean squared error relative to the true effect.

Table E.1 reports the bias, standard deviation ('sd'), root mean squared error

('rmse'), true effect ('true'), and the relative root mean squared error in percent of the true effect ('relr') of the respective estimators of θ_1^{n0}, θ_1^{n1}, Δ_1^{ap}, $\theta_1^{ap}(1)$, $\theta_1^{ap}(0)$, $\delta_1^{ap}(1)$, and $\delta_1^{ap}(0)$ for the linear model. In this case, the identifying assumptions underlying both the CiC (Panel A.) and DiD (Panel B.) estimators are satisfied. Specifically, the homogeneous time trend on the cross-sectional observation unit satisfies any of the common trend assumptions in Deuchert, Huber, and Schelker (2019), while the monotonicity of Y in U and the independence of T and U satisfies the key assumptions of this paper. For this reason any of the estimates in Table E.1 are close to being unbiased and appear to converge to the true effect at the parametric rate when comparing the results for the two different sample sizes.[1]

Table E.2 provides the results for the exponential outcome model, in which the time trend is heterogeneous and interacts with U through the nonlinear link function. While the CiC assumptions hold (Panel A.), average time trends are heterogeneous across complier types such that the DiD approach (Panel B.) of Deuchert, Huber, and Schelker (2019) is inconsistent. Accordingly, the biases of the CiC estimates generally approach zero as the sample size increases, while this is not the case for the DiD estimates. CiC yields a lower root mean squared error than the respective DiD estimator in all but one case (namely $\hat{\delta}_1^{ap}(0)$ with $N = 1,000$) and its relative attractiveness increases in the sample size due to its lower bias.[2]

In our next simulation design, we maintain the exponential outcome model but assume D to be selective w.r.t. U rather than random. To this end, the treatment model in (E) is replaced by $D = I\{U + Q > 0\}$, with the independent variable $Q \sim N(0,1)$ being an unobserved term. The average of U among the treated and no-treated amounts to 0.24 and -0.24, respectively. This treatment selectivity entails

[1] In contrast, two stage least squares regression using D as instrument for M is inconsistent due to the direct effects violating the IV exclusion restriction. The IV estimate neither recovers Δ_1^{ap}, nor $\delta_1^{ap}(1)$, nor $\delta_1^{ap}(0)$, with the bias amounting to approximately 4, 5, and 6, respectively, for the three parameters with the sample sizes considered. This motivates the application of our method to verify the IV exclusion restriction in Section 4.

[2] However, we can easily modify the DGP underlying Table E.2 to match a scenario in which also CiC is inconsistent, e.g. by a violation of Assumption 3. For instance, when changing the distribution of U to $U|T = 0 \sim Unif(-1, 1)$ and $U|T = 1 \sim Unif(0, 1)$ such that it depends on T, we obtain non-negligible biases in the CiC estimates that do not vanish as the sample size increases.

Table E.2: Nonlinear model with random treatment

	$\hat{\theta}_1^{n0}$	$\hat{\theta}_1^{n1}$	$\hat{\Delta}_1^{ap}$	$\hat{\theta}_1^{ap}(1)$	$\hat{\theta}_1^{ap}(0)$	$\hat{\delta}_1^{ap}(1)$	$\hat{\delta}_1^{ap}(0)$
A. Changes-in-Changes							
				N=1,000			
bias	0.01	-0.14	-0.48	-0.35	-0.11	-0.37	-0.13
sd	0.48	5.08	8.47	6.20	1.16	8.64	4.23
rmse	0.48	5.08	8.48	6.21	1.17	8.65	4.23
true	3.49	68.09	52.42	47.70	4.72	47.70	4.72
relr	0.14	0.07	0.16	0.13	0.25	0.18	0.90
				N=4,000			
bias	-0.01	0.01	-0.00	-0.11	-0.07	0.07	0.11
sd	0.25	2.63	4.37	3.20	0.66	4.44	2.04
rmse	0.25	2.63	4.37	3.20	0.66	4.44	2.04
true	3.49	68.09	52.45	47.73	4.72	47.73	4.72
relr	0.07	0.04	0.08	0.07	0.14	0.09	0.43
B. Difference-in-Differences							
				N=1,000			
bias	-0.27	-8.91	14.42	11.46	-1.49	15.91	2.96
sd	0.46	2.62	2.58	2.62	0.47	2.61	0.47
rmse	0.53	9.29	14.65	11.76	1.56	16.12	2.99
true	3.49	68.09	52.42	47.70	4.72	47.70	4.72
relr	0.15	0.14	0.28	0.25	0.33	0.34	0.63
				N=4,000			
bias	-0.28	-8.79	14.51	11.57	-1.51	16.02	2.94
sd	0.24	1.28	1.26	1.28	0.25	1.27	0.23
rmse	0.37	8.88	14.57	11.64	1.53	16.07	2.95
true	3.49	68.09	52.45	47.73	4.72	47.73	4.72
relr	0.11	0.13	0.28	0.24	0.32	0.34	0.62

Note: 'bias', 'sd', and 'rmse' provide the bias, standard deviation, and root mean squared error of the respective estimator. 'true' and 'relr' are the respective true effect as well as the root mean squared error relative to the true effect.

non-negligible differences in mean potential outcomes across treatment groups, e.g. $E[Y_1(1,1)|D=1] - E[Y_1(1,1)|D=0] = 29.1$. Under this violation of Assumption 5, the shares and effects of affected positively are no longer identified, which is confirmed by the simulation results presented in Table E.3. The bias in the CiC based total, direct, and indirect effects on affected positively do not vanish as the sample size increases. Furthermore, under non-random assignment of D (while maintaining monotonicity of M in D), the not-affected at 0 and 1 respective distributions of U differ across treatment. Therefore, average direct effects among the total of not-

Table E.3: Nonlinear model with non-random treatment

	$\hat{\theta}_1^{0,1}(1)$	$\hat{\theta}_1^{1,0}(0)$	$\hat{\Delta}_1^{ap}$	$\hat{\theta}_1^{ap}(1)$	$\hat{\theta}_1^{ap}(0)$	$\hat{\delta}_1^{ap}(1)$	$\hat{\delta}_1^{ap}(0)$
A. Changes-in-Changes							
				N=1,000			
bias	0.02	0.13	47.21	40.19	-1.44	48.64	7.02
sd	0.71	4.56	5.45	4.11	0.75	5.53	2.92
rmse	0.71	4.56	47.52	40.40	1.62	48.96	7.60
true	4.41	54.19	52.42	47.70	4.72	47.70	4.72
relr	0.16	0.08	0.91	0.85	0.34	1.03	1.61
				N=4,000			
bias	-0.00	0.06	47.38	40.13	-1.53	48.91	7.25
sd	0.38	2.35	2.84	2.04	0.38	2.86	1.51
rmse	0.38	2.35	47.47	40.18	1.57	48.99	7.40
true	4.40	54.18	52.45	47.73	4.72	47.73	4.72
relr	0.09	0.04	0.90	0.84	0.33	1.03	1.57
B. Difference-in-Differences							
				N=1,000			
bias	0.35	19.98	29.00	27.65	0.04	28.96	1.35
sd	0.67	2.48	2.46	2.48	0.67	2.51	0.45
rmse	0.75	20.14	29.11	27.76	0.67	29.07	1.43
true	4.41	54.19	52.42	47.70	4.72	47.70	4.72
relr	0.17	0.37	0.56	0.58	0.14	0.61	0.30
				N=4,000			
bias	0.34	20.02	28.98	27.65	0.02	28.96	1.33
sd	0.35	1.22	1.19	1.22	0.35	1.24	0.23
rmse	0.49	20.06	29.01	27.68	0.35	28.99	1.35
true	4.40	54.18	52.45	47.73	4.72	47.73	4.72
relr	0.11	0.37	0.55	0.58	0.07	0.61	0.29

Note: 'bias', 'sd', and 'rmse' provide the bias, standard deviation, and root mean squared error of the respective estimator. 'true' and 'relr' are the respective true effect as well as the root mean squared error relative to the true effect.

affected at 0 or 1, respectively, are not identified. Yet, $\theta_1^{1,0}(1)$, which is still identified by the same estimator as before, yields the direct effect among treated not-affected at 0 (as affected negatively do not exist). Likewise, $\theta_1^{0,1}(0)$ corresponds to the direct effect on non-treated not-affected at 1. Indeed, the results in Table E.3 suggest that both parameters are consistently estimated by the CiC approach (Panel A.).

Finally, we also consider a violation of Assumption 6 by relaxing monotonicity of M in D. We do so by modifying the mediator equation to $M = I\{(2\kappa - 1) \cdot D + U + V > 0\}$, with $\kappa \sim Binom(0.2)$ being a randomly and binomially distributed variable,

implying that the coefficient on D is either 1 or -1 with a probability of 80% or 20%, respectively. This entails a defier share of roughly 9% in the population, while we otherwise maintain the specification underlying the results in Table E.3. We note that $\theta_1^{1,0}(1)$ now corresponds to the direct effect on treated not-affected at 0 and affected negatively, $\theta_1^{0,1}(0)$ on non-treated not-affected at 1 and affected negatively. Table E.4 provides the results. Again, CiC performs decently for estimating $\theta_1^{1,0}(1)$ and $\theta_1^{0,1}(0)$ as suggested by Theorem 1, while non-negligible relative root mean squared error arise for the remaining parameters.

Table E.4: Nonlinear model with non-random treatment and non-monotonicity

	$\hat{\theta}_1^{0,1}(1)$	$\hat{\theta}_1^{1,0}(0)$	$\hat{\Delta}_1^{ap}$	$\hat{\theta}_1^{ap}(1)$	$\hat{\theta}_1^{ap}(0)$	$\hat{\delta}_1^{ap}(1)$	$\hat{\delta}_1^{ap}(0)$
A. Changes-in-Changes							
			N=1,000				
bias	0.06	0.24	65.65	55.29	-3.76	69.41	10.35
stdev	0.62	4.90	10.98	7.74	0.86	11.25	6.47
rmse	0.62	4.91	66.56	55.83	3.86	70.31	12.21
true	5.62	54.19	52.45	47.73	4.72	47.73	4.72
relr	0.11	0.09	1.27	1.17	0.82	1.47	2.59
			N=4,000				
bias	0.02	0.10	65.91	55.01	-3.84	69.75	10.90
stdev	0.31	2.49	5.80	4.03	0.46	5.99	3.23
rmse	0.32	2.49	66.17	55.16	3.86	70.00	11.36
true	5.63	54.18	52.45	47.73	4.72	47.73	4.72
relr	0.06	0.05	1.26	1.16	0.82	1.47	2.41
B. Difference-in-Differences							
			N=1,000				
bias	0.79	21.78	31.59	30.24	1.70	29.90	1.36
stdev	0.54	2.82	2.79	2.81	0.56	2.82	0.46
rmse	0.96	21.97	31.72	30.37	1.79	30.03	1.43
true	5.62	54.19	52.45	47.73	4.72	47.73	4.72
relr	0.17	0.41	0.60	0.64	0.38	0.63	0.30
			N=4,000				
bias	0.80	21.76	31.54	30.21	1.70	29.84	1.33
stdev	0.27	1.36	1.33	1.36	0.28	1.35	0.24
rmse	0.84	21.80	31.57	30.24	1.72	29.87	1.35
true	5.63	54.18	52.45	47.73	4.72	47.73	4.72
relr	0.15	0.40	0.60	0.63	0.37	0.63	0.29

Note: 'bias', 'sd', and 'rmse' provide the bias, standard deviation, and root mean squared error of the respective estimator. 'true' and 'relr' are the respective true effect as well as the root mean squared error relative to the true effect.

F Background Information for Applications

F.1 JOBS II Evaluation

The JOBS II was a modified version of the earlier JOBS programme, which had been found to improve labour market outcomes such as job satisfaction, motivation, earnings, and job stability, see Caplan, Vinokur, Price, and van Ryn (1989) and Vinokur, van Ryn, Gramlich, and Price (1991), as well as mental health, see Vinokur, Price, and Caplan (1991). According to the results of Vinokur, Price, and Schul (1995), the JOBS II programme increased re-employment rates and improved mental health outcomes, especially for participants having an elevated risk of depression. The JOBS interventions had an important impact in the academic literature (see e.g. Wanberg, 2012, Liu, Huang, and Wang, 2014) and the methodology was implemented in field experiments in Finland (Vuori, Silvonen, Vinokur, and Price, 2002, Vuori and Silvonen, 2005) and the Netherlands (Brenninkmeijer and Blonk, 2011), suggesting positive effects on labour market integration in either case. Imai, Keele, and Tingley (2010) analyse Jobs II in a mediation context as well, but consider a different mediator, namely job search self-efficacy, and a different identification strategy based on selection on observables.

In the JOBS II intervention, individuals responded to a screening questionnaire that collected pre-treatment information on mental health in the baseline period. Based on the latter, individuals were classified as having either a high or low depression risk and those with a high risk were oversampled before the training was randomly assigned. Randomization was followed by yet another questionnaire sent out two weeks before the actual job training, see Vinokur, Price, and Schul (1995), which also provided information on whether an individual had been assigned the training. Consequently, the data collected in that questionnaire must be considered post-treatment as they could be affected by learning the assignment. Therefore, we rely on the earlier screening data as the relevant pre-treatment period prior to random programme assignment.

The job training consisted of five 4-hours seminars conducted in morning sessions during one week between March 1 and August 7, 1991. Members of the treatment group who participated in at least four of the five sessions received USD 20. Each of the standardized training sessions consisted, among other aspects, of the learning and practicing of job search and problem-solving skills. The control group received a booklet with information on job search methods (Vinokur, Price, and Schul, 1995, p. 44-49).

F.2 Paid Maternal Leave Reform

There is a large literature on the impact of maternal or parental leave on female labour supply, earnings, or fertility, see for instance Lalive and Zweimüller (2009), Lalive, Schlosser, Steinhauer, and Zweimüller (2014), Fitzenberger, Steffes, and Strittmatter (2016), Byker (2016), Dahl, Løken, Mogstad, and Salvanes (2016), Olivetti and Petrongolo (2017), and Zimmert and Zimmert (2020). The design of maternal or paternal leave programs varies substantially across countries and estimation results depend heavily on the design of such programs with respect to the leave duration, the income replacement rate, job protection regulation, the availability of paid leave to either parent, etc. (Olivetti and Petrongolo, 2017).

In Switzerland, paid maternal leave was only introduced in 2005. Before, the Law on Manufacturing of 1877 just prohibited maternal labour supply for 8 weeks, with at least 6 weeks taken right after childbirth. In 1945 the constitutional bases for a paid maternal leave were established. However, numerous attempts to actually introduce paid maternal leave were all rejected in nation-wide popular ballots, the last unsuccessful attempt only dating back to 1999. Finally, on September 24, 2004, a majority of 55.4% of Swiss citizens voted in favour of the introduction of 14 weeks of paid maternal leave, with a replacement rate of 80% and a cap at CHF 172 per day in 2005. Paid maternal leave is covered through the Swiss fund for loss of earnings and maternal pay. The reform took effect on July 1, 2005. Job protection regulation remained unaffected and protection against dismissal lasts during the

entire pregnancy and 16 weeks after childbirth.

The political campaigning and discussions on the various topics on the agenda (there were four federal propositions in September) typically start two to three months before. Given that all previous attempts to introduce paid maternal leave were rejected in popular ballots, the latest in 1999, and that the subsequent acceptance with 55.4% was far from overwhelming, important anticipation effects are fairly unlikely. The post-treatment period contains information from the 2007 questionnaire. We do not use data from 2005 and 2006 because interviews of the Swiss Labour Force Survey are only conducted up to the end of June each year. This makes 2005 a pre-treatment period and childbirth in early 2006 is the result of fertility decisions before or just around the introduction of paid maternal leave legislation in July 2005.

References

Brenninkmeijer, V., and R. W. Blonk (2011): "The Effectiveness of the JOBS Program Among the Long-Term Unemployed: A Randomized Experiment in the Netherlands," *Health Promotion International*, 27, 220–229.

Byker, T. (2016): "Paid Parental Leave Laws in the United States: Does Short-Duration Leave Affect Women's Labor-Force Attachment?," *American Economic Review*, 106, 242–262.

Callaway, B. (2016): "Quantile Treatment Effects in R: The qte Package," *Working Paper*.

Caplan, R. D., A. D. Vinokur, R. H. Price, and M. van Ryn (1989): "Job Seeking, Reemployment, and Mental Health: A Randomized Field Experiment in Coping with Job Loss," *Journal of Applied Psychology*, 74, 759–769.

Dahl, G. B., K. V. Løken, M. Mogstad, and K. V. Salvanes (2016): "What

is the Case for Paid Maternity Leave?," *Review of Economics and Statistics*, 98, 655–670.

Deuchert, E., M. Huber, and M. Schelker (2019): "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*, 37, 710–720.

Fitzenberger, B., S. Steffes, and A. Strittmatter (2016): "Return-to-Job During and After Maternity Leave," *International Journal of Human Resource Management*, 27(2), 803–831.

Imai, K., L. Keele, and D. Tingley (2010): "A General Approach to Causal Mediation Analysis," *Psychological Methods*, 15, 309–334.

Lalive, R., A. Schlosser, A. Steinhauer, and J. Zweimüller (2014): "Parental Leave and Mothers' Careers: The Relative Importance of Job Protection and Cash Benefits," *Review of Economic Studies*, 81, 219–265.

Lalive, R., and J. Zweimüller (2009): "How Does Parental Leave Affect Fertility and Return to Work? Evidence from Two Natural Experiments," *Quarterly Journal of Economics*, 124, 1363–1402.

Liu, S., J. L. Huang, and M. Wang (2014): "Effectiveness of Job Search Interventions: A Meta-Analytic Review," *Psychological Bulletin*, 140, 1009–1041.

Olivetti, C., and B. Petrongolo (2017): "The Economic Consequences of Family Policies: Lessons from a Century of Legislation in High-Income Countries," *Journal of Economic Perspectives*, 31, 205–230.

Vinokur, A. D., R. H. Price, and R. D. Caplan (1991): "From Field Experiments to Program Implementation: Assessing the Potential Outcomes of an Experimental Intervention Program for Unemployed Persons," *American Journal of Community Psychology*, 19, 543–562.

Vinokur, A. D., R. H. Price, and Y. Schul (1995): "Impact of the JOBS Intervention on Unemployed Workers Varying in Risk for Depression," *American Journal of Community Psychology*, 23, 39–74.

Vinokur, A. D., M. van Ryn, E. M. Gramlich, and R. H. Price (1991): "From field experiments to program implementation: Assessing the potential outcomes of an experimental intervention program for unemployed persons," *Journal of Applied Psychology*, 76, 213–219.

Vuori, J., and J. Silvonen (2005): "The Benefits of a Ppreventive Job Search Program on Re-Employment and Mental Health at 2-Year Follow-Up," *Journal of Occupational and Organizational Psychology*, 78, 43–52.

Vuori, J., J. Silvonen, A. D. Vinokur, and R. H. Price (2002): "The Työhön Job Search Program in Finland: Benefits for the Unemployed With Risk of Depression or Discouragement," *Journal of Occupational Health Psychology*, 7, 5–19.

Wanberg, C. R. (2012): "The Individual Experience of Unemployment," *Annual Review of Psychology*, 63, 369–396.

Zimmert, F., and M. Zimmert (2020): "Paid parental leave and maternal reemployment: Do part-time subsidies help or harm?," *Working Paper*.