A Framework for Separating Individual Treatment Effects From Spillover, Interaction, and General Equilibrium Effects

Martin Huber (University of Fribourg)
Andreas Steinmayr (University of Munich)

Discussion Paper No. 21

March 23, 2017

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

A framework for separating individual treatment effects from spillover, interaction, and general equilibrium effects

Martin Huber* and Andreas Steinmayr**

*University of Fribourg, **University of Munich, IfW Kiel, and IZA

Abstract: This paper suggests a causal framework for disentangling individual level treatment effects and interference effects, i.e., general equilibrium, spillover, or interaction effects related to treatment distribution. Thus, the framework allows for a relaxation of the Stable Unit Treatment Value Assumption (SUTVA), which assumes away any form of treatment-dependent interference between study participants. Instead, we permit interference effects within aggregate units, for example, regions or local labor markets, but need to rule out interference effects between these aggregate units. Borrowing notation from the causal mediation literature, we define a range of policy-relevant effects and formally discuss identification based on randomization, selection on observables, and difference-in-differences. We also present an application to a policy intervention extending unemployment benefit durations in selected regions of Austria that arguably affected ineligibles in treated regions through general equilibrium effects in local labor markets.

Keywords: treatment effect, general equilibrium effects, spillover effects, interaction effects, interference effects, inverse probability weighting, propensity score, mediation analysis, difference-in-differences.

JEL classification: C21, C31.

Addresses for correspondence: Martin Huber, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland, (martin.huber@unifr.ch); Andreas Steinmayr, University of Munich (LMU), Ludwigstrasse 33, 80539 Munich, Germany, and IfW Kiel, IZA (andreas.steinmayr@econ.lmu.de).

Financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 is gratefully acknowledged. We thank Josef Zweimüller for his support with the empirical application. Joachim Winter provided helpful comments on the draft.

1 Introduction

Most studies on treatment evaluation either implicitly or explicitly rule out general equilibrium, spill-over, or interaction effects related to individual treatment assignment. The Stable Unit Treatment Value Assumption (SUTVA), see for instance Rubin (1990), formalizes the absence of any such interference effects between study participants. However, the satisfaction of SUTVA appears unrealistic in many scenarios including labor market, development, and educational interventions, see Heckman, Lochner, and Taber (1998) for a critical discussion. Considering for instance a training program, the share of individuals who receive some training in a region may have an impact on someone's employment probability even net of the individual training status, due to an increase in the regional supply of a particular skill. When assessing the effects of book provision to high school students, spillover effects may occur through sharing the books with peers in class who did not receive the books. In both examples, the overall treatment effect would be different from the (average) individual one, see Sobel (2006) for a framework for characterizing the bias in the presence of interference. Interference effects are also a likely reason why many interventions that have been deemed successful in a small-scale randomized control trial (RCT) fail to produce similar effects when scaled-up to a larger population, see for example Deaton and Cartwright (2016) for a discussion of this problem.

Drawing from the literature on causal mediation analysis, this paper proposes a general and nonparametric framework for separating individual level treatment effects from interference effects. After defining a range of policy-relevant parameters, we systematically discuss which effects of interest can be identified under particular randomization, selection on observables, and difference-in-differences assumptions. One crucial condition underlying all our approaches is that SUTVA has to be fulfilled on some aggregate, e.g. regional or group level, while it (in contrast to the standard literature) may be violated on the individual level. Throughout the paper, we will refer to the entities at the aggregate level as regions. Regional SUTVA allows for interference effects between the individuals within regions, but rules out such effects across regions. Given regional SUTVA, the total treatment effect may be split up into two causal mechanisms: (i) an individual effect and (ii) a within-region interference effect net of the individual impact that is

driven by the treatment assignment of other individuals in the region.

The regional treatment may be defined as a binary variable, e.g. whether a region is targeted by a treatment at all or not, or by a multivalued regional treatment intensity, reflecting, for instance, the proportion of individuals receiving a particular treatment in that region. The individual treatment is a binary indicator for whether a particular individual is treated (note that even in targeted regions, only a subgroup may actually be treated). As an important feature of our framework, individual and regional treatment effects may interact arbitrarily. This permits that interference effects depend on the individual treatment status and that individual treatment effects depend on the regional treatment intensity. Albeit this makes the analysis more complex, it appears important in practice, as for instance, an individual training may be more effective in a labor market where only few other individuals obtain a similar skill.

As one conceptual contribution, we distinguish between "natural" effects (in the denomination of Pearl (2001)) defined upon the potential individual treatment states that would occur under a particular regional treatment intensity, and "controlled" effects, where individual treatments are forced to take a specific value. This distinction is important, as natural effects are per definition consistent with a particular regional treatment intensity, while controlled effects may refer to practically non-realizable causal comparisons. An example of the latter kind is the average individual treatment effect in the total population (comparing the potential outcomes under 100% and 0% treatment), under the condition that only 50% of this population are treated. If only half of the population can be treated, a parameter that is based on a comparison on individually treating all vs. no-one does not reflect a practically feasible policy choice.

As an empirical contribution, we reconsider data from Lalive, Landais, and Zweimüller (2015) who study the spillover effects of a large-scale extension of unemployment benefits in selected regions of Austria and find that this policy decreased the job-search duration of ineligible individuals in treated regions. We use our framework to provide a sharper definition of the identified effects and apply our difference-in-differences methodology to identify total effects on eligibles and spillover effects on ineligibles. Furthermore, we compare estimation based on (i) a common trend assumption within groups having the same eligibility status and (ii) a stronger common

trend assumption across groups and discuss the testable implications of the latter. The results suggest that the stronger common trend assumption, as considered in some estimations of Lalive, Landais, and Zweimüller (2015), is most likely violated. Even though the estimated effects under common trends within groups are smaller in magnitude, they qualitatively confirm the finding under the stronger assumption, namely a substantial positive effect on eligibles and a negative spillover effect on ineligibles.

Our paper is distinct from conventional approaches in the literature on spillover and peer effects that typically rely on structural assumptions not imposed here. See for instance Graham (2008), who shows that specific conditional variance restrictions on outcomes entail point identification of peer effects if outcomes are linear in average characteristics within some region or group, as discussed in Manski (1993). Our approach of defining and identifying effects is more closely related to nonparametric mediation analysis, which aims at disentangling the causal mechanisms through which an intervention affects an outcome of interest. See for instance Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), VanderWeele (2009), Hong (2010), Imai, Keele, and Yamamoto (2010), and Huber (2014), among many others. Our framework is (at least in terms of notation) also related to the dynamic treatment effects literature aiming to analyze sequences of treatments, see for instance Robins (1986, 1989), Robins, Hernan, and Brumback (2000), Lechner (2009), and Lechner and Miquel (2010), which, however, does not consider natural effects. We also make use of the principal stratification framework of Frangakis and Rubin (2002) to consider effects for specific subpopulations that are defined by the relation between the individual treatment state and a particular regional treatment intensity. Principal stratification has been considered in the context of mediation analysis for instance by Rubin (2004) and VanderWeele (2008, 2012)) and in the context of spillover effects by Forastiere, Mealli, and VanderWeele (2016).

Also Angelucci and Di Maro (2015) discuss the identification of interference effects under different identifying strategies based on experiments and non- and quasi-experimental methods such as conditional independence, regression discontinuity, and instrumental variable assumptions. However, they solely discuss the identification of the total effect on eligibles (i.e.

individually treated) in treated regions as well as the indirect effect on ineligibles in treated areas, while we consider a broader set of effects. Ferracci, Jolivet, and van den Berg (2010) is a further contribution in this context, who similarly to our study explicitly invoke a regional SUTVA[1] and consider a selection on observables assumption about the regional and individual treatment assignments. They estimate the mean potential outcomes when either every or no subject is individually treated as a function of the regional treatment intensity. While this is suitable for detecting violations of the individual SUTVA (namely if the mean potential outcomes are heterogeneous across the intensity), the mean potential outcomes investigated cannot exist in reality unless regional treatment is either 100% or 0%. Our paper differs as we also aim for practically feasible parameters, in particular natural effects, based on alternative sets of identifying assumptions.

The remainder of this paper is organized as follows. Section 2 defines and discusses interference and individual treatment effects for various populations. Section 3 provides identification results under various sets of assumptions related to randomization, selection on observables, and difference-in-differences. Section 4 presents an applications to a labor market intervention in Austria that extended unemployment benefits in selected regions. Section 5 concludes.

2 Definition of effects

2.1 A general framework for individual and interference effects

We denote by Z the regional treatment intensity, by D the individual treatment assignment, and by Y an individual level outcome of interest. Importantly, Z might affect Y also other than through its effect on the individual treatment decision D, reflecting general equilibrium, interaction, or other interference effects. The regional intensity of training for job seekers, for instance, may affect the employment probability net of individual training participation through general equilibrium effects: The larger the proportion of trained individuals, the lower may be

[1]See for instance Hong and Raudenbush (2006) and Forastiere, Mealli, and VanderWeele (2016) for further studies invoking SUTVA on an aggregate level to identify interference effects. Related assumptions are typically also made in the 'structural' literature on peer effects and spillovers, see for instance the the 'no cross neighborhood spillovers' restriction in Graham, Imbens, and Ridder (2010).

the employment chances for some or all individuals in the labor market, conditional on their own treatment state. As a second example, let Z be a binary indicator for the availability of income support for disadvantaged households in a subset of regions. D reflects whether a household is eligible to income support and is zero whenever $Z = 0$. In treated regions where $Z = 1$, households below a particular poverty index threshold benefit from income support ($D = 1$), while wealthier households are ineligible ($D = 0$). However, even the latter may benefit from spillovers through increased consumption and purchasing power of eligible households.

The individual treatment is assumed to be binary ($D \in \{1, 0\}$), participation vs. non-participation.[2] Depending on the application, Z might be either binary or multivalued to reflect the regional intensity of the program, e.g. the share of a region's relevant population the treatment is offered to.[3] To define the parameters of interest, we make use of the potential outcome notation, see for instance Rubin (1974), and denote by $Y(z, d)$ the potential outcome under the regional treatment (intensity) z and the individual treatment d. Furthermore, we denote by T some target population of interest, which may, for instance, comprise all individuals receiving the individual treatment ($D = 1$). This allows defining average individual and interference effects for some population $T = t$ respectively:

$$\begin{aligned} \delta_t(z) &= E[Y(z,1) - Y(z,0)|T = t] \text{ with } z \text{ in the support of } Z, \\ \theta_t(z', z, d) &= E[Y(z', d) - Y(z, d)|T = t] \text{ with } z' \neq z \text{ and } z', z \text{ in the support of } Z \text{ and } d \in \{0, 1\}. \end{aligned} \tag{1}$$

$\delta_t(z)$ reflects the impact of the individual treatment D given a particular regional treatment intensity z. The average individual treatment effect is therefore allowed to be a function of the regional treatment intensity. For instance, an individual training could be less effective if a larger share of labor market participants receive the same qualification. $\theta_t(z', z, d)$ corresponds to the interference effect when comparing the regional treatment intensities z' vs. z, given a particular

[2]The framework could also be extended to multivalued individual treatments, which is omitted for simplicity.

[3]Even though we refer to Z as intensity throughout the paper, which suggests considering different proportions of treated individuals across regions, we would like to point out that different values in Z may more generally reflect different distributions of treated individuals.

individual treatment d. The interference effect may therefore be a function of the individual treatment state. For instance, the spillover effect of providing some students with books on academic performance (see for instance Frölich and Michaelowa (2011)) may be larger for students not receiving books ($D = 0$) than for students receiving books ($D = 1$), if the individual treatment effect partly crowds out the spillover effect. Only if there are no interaction effects between Z and D on Y are $\delta_t(z)$ and $\theta_t(z', z, d)$ not functions of z and d, respectively, and may be written as δ_t and $\theta_t(z', z)$.[4] We will henceforth not make this restriction and allow for interactions between Z and D. Furthermore, note that if the regional treatment intensity has only two levels, the interference effect reduces to a binary comparison of $\theta_t(d) = E[Y(1, d) - Y(0, d)|T = t]$ with $z = 1$ and $z = 0$ denoting the higher and lower regional treatment intensity, respectively. For the sake of easy of notation, we stick to the binary case for most of the remainder of this paper, but deviate whenever appropriate (as in parts of Sections 2.2 and 3).

We introduce some further notation that will be helpful for defining interesting target populations determined by how the individual treatment state varies with the regional treatment intensity. Denote by $D_i(z)$ the potential individual treatment of any subject i in the population when setting the regional treatment to $Z = z$. This notation is motivated by the presumption that individual treatment status is a function of program availability in the region. Similar to the principal stratification approach of Frangakis and Rubin (2002) and the instrumental variable framework of Angrist, Imbens, and Rubin (1996), any individual belongs to one of four compliance types $\mathcal{T}$, defined by the potential individual treatment states under $z = 1$ and $z = 0$: always takers ($\mathcal{T}_i = a : D_i(1) = D_i(0) = 1$) who are individually treated both under high and low regional treatment intensity, compliers ($\mathcal{T}_i = c : D_i(1) = 1, D_i(0) = 0$) who get individual treatment under high, but not under low regional treatment intensity, defiers ($\mathcal{T}_i = d : D_i(1) = 0, D_i(0) = 1$) who behave opposite to the compliers, and never takers ($\mathcal{T}_i : D_i(1) = D_i(0) = 0$) who do not receive individual treatment under either regional treatment intensity. $\mathcal{T}_i$ can not be pinned down for any individual without further assumptions, because either $D_i(1)$ or $D_i(0)$ is observed (depending on whether Z_i is one or zero), but never both.

[4] This is satisfied under the constant unit-level treatment effect assumption of Robins (2003) (requiring that $Y_i(1, 1) - Y_i(0, 1) = Y_i(1, 0) - Y_i(0, 0)$ for any individual i).

2.2 Two empirical examples

In this subsection, we consider two empirical examples to demonstrate which kind of effects have been typically analyzed in the literature on interference effects.[5] Our first example is PROGRESA, a conditional cash transfer program for poor households in Mexico, in which treatment villages are randomly chosen, but only parts of the households in treated villages are actually offered the cash transfer as a function of a poverty index.[6] The lower treatment intensity ($Z = 0$) corresponds to zero such that no individual is treated. That is, $\Pr(D = 1|Z = 0) = 0$ and $D_i(0) = 0$ for all i, which rules out defiers and always takers (unless people people moved across regions, which would, however, violate regional SUTVA outlined in Section 3). In treated villages ($Z = 1$), households below a particular poverty threshold were entitled to cash transfers ($D = 1$), while wealthier households were not ($D = 0$). Therefore, the types have a clear and policy-relevant interpretation: Never takers are noneligible, wealthier households, while compliers are poorer and eligible if the village is randomized in.

The design of PROGRESA allows identifying the spillover effect on the never takers in the absence of the individual treatment, $\theta_n(0)$, under the assumption that the stable unit treatment valuation assumption (SUTVA) holds on the regional level as formalized in Section 3. Because (i) Z is random, (ii) type $\mathcal{T}_i$ of any individual i is a deterministic function of the observed poverty index, and (iii) D is a deterministic function of Z and the poverty index. Therefore, never takers are identified in both treated and non-treated villages, such that $\theta_n(0) = E[Y(1,0) - Y(0,0)|\mathcal{T} = n] = E[Y|Z = 1, \mathcal{T} = n] - E[Y|Z = 0, \mathcal{T} = n]$, see also Section 3.1. Angelucci and Giorgi (2009) and others exploit this strategy and even though the authors utilize a somewhat different notation, their "indirect treatment effect" corresponds to $\theta_n(0)$.[7]

[5] Our empirical application based on data from Lalive, Landais, and Zweimüller (2015) provides yet another example and is discussed in Section 4.

[6] Therefore, PROGRESA is a so-called partial-population experiment, see Moffitt (2001). Further examples for partial-population (quasi-)experiments include Miguel and Kremer (2004), who study the spillover effects of a deworming treatment that was randomized among schools in Kenya, Baird, Bohren, McIntosh, and Ozler (2012), who assess the spillover effects of a cash transfer program in Malawi, Dahl, Løken, and Mogstad (2014), who estimate the peer effects of paid paternity leave (on the peers' brothers and coworkers) in Norway using a regression discontinuity design.

[7] Angelucci and Giorgi (2009) find that PROGRESA cash transfers to eligible households indirectly increase the consumption of ineligible households in the same villages, while Bobonis and Finan (2009) use variation in the school participation of program-eligible children to identify peer effects on the schooling decisions of ineligible children. Lalive and Cattaneo (2009) use information on a child's eligible classroom peers in treated villages as

Furthermore, Angelucci and Giorgi (2009) also estimate the (total) average effect of the policy intervention on the compliers (eligible households): $\Delta_c = E[Y(1,1) - Y(0,0)|\mathcal{T} = c] = E[Y|Z = 1, \mathcal{T} = c] - E[Y|Z = 0, \mathcal{T} = c]$. The latter parameter comprises both the individual and interference effects on eligible households:

$$\begin{aligned} \Delta_c &= E[Y(1,1) - Y(1,0)|\mathcal{T} = c] + E[Y(1,0) - Y(0,0)|\mathcal{T} = c] = \delta_c(1) + \theta_c(0) \\ &= E[Y(0,1) - Y(0,0)|\mathcal{T} = c] + E[Y(1,1) - Y(0,1)|\mathcal{T} = c] = \delta_c(0) + \theta_c(1). \end{aligned} \tag{2}$$

(2) shows that the total effect on the compliers adds up to the individual treatment effect in villages receiving cash transfers ($\delta_c(1)$) and the interference effect when not treated individually ($\theta_c(0)$). Alternatively, it adds up to the individual treatment effect in villages not receiving cash transfers ($\delta_c(0)$) and the interference effect when treated individually ($\theta_c(1)$). That is, the two decompositions differ with respect to whether the interaction effects of Z and D are assigned to the individual or to the interference effect. Arguably, $\delta_c(0)$ appears particularly interesting, because it corresponds to the individual effect if no one else received the treatment in the the region. This corresponds to the effect we have in mind when imposing the individual level SUTVA, which rules out any interference effects. However, whenever $Z = 0$ represents a regional treatment intensity of zero as in PROGRESA, $\delta_c(0)$ as well as $\theta_c(1)$ cannot be nonparametrically identified. The reason is that $\Pr(D = 1|Z = 0) = 0$, such that $E[Y(0,1)|T = t]$ cannot be inferred for any t.[8]

As a second empirical example, Crépon, Duflo, Gurgand, Rathelot, and Zamora (2012) assess a randomized job placement assistance program in France, where the probability to receive the program differs across regions, which corresponds to a multivalued Z. They find that the regional intensity of the program negatively affects the employment chances of individuals not taking the treatment. The analysis differs from PROGRESA in that not only Z, but also the individual treatment assignment D is randomized.[9] This implies that asymptotically, characteristics and

an instrument for peer group schooling. This instrument varies within villages, allowing identifying peer effects within rather than across villages. They also decompose individual and spillover effects for eligible students based on various structural assumptions.

[8]Lalive and Cattaneo (2009) nevertheless decompose the individual and interference effects among compliers, but require a more tightly specified structural model that parametrically determines how (and through which channels) interference effects come about.

[9]See Angelucci, Prina, Royer, and Samek (2015) for a further study relying on double randomization and Hudgens and Halloran (2008) for a formal discussion of identification and inference in such a context.

effects do not vary across various populations T defined in terms of Z, D, or $\mathcal{T}$ and correspond to the total population:[10] $\delta_t(z) = \delta(z) = E[Y(z,1) - Y(z,0)]$ and $\theta_t(z', z, d) = \theta(z', z, d) = E[Y(z', d) - Y(z, d)]$. The authors consider four regional treatment intensities corresponding to the share of treated job seekers (with 0 corresponding to 0% and 1 to 100%): $z \in \{0, 0.25, 0.5, 0.75\}$. Again, $\delta(0)$ is not identified nonparametrically, but the authors assess effect heterogeneity in $\delta(z)$ over positive values in z. This allows partly identifying interaction effects between Z and D over such values z. To see this, consider the following saturated (and thus nonparametric) potential outcome model:

$$\begin{aligned} E[Y(Z,D)] &= \beta_0 + \beta_1 D + \beta_2 I\{Z = 0.25\} + \beta_3 I\{Z = 0.5\} + \beta_4 I\{Z = 0.75\} \\ &+ \beta_5 DI\{Z = 0.25\} + \beta_6 DI\{Z = 0.5\} + \beta_7 DI\{Z = 0.75\}. \end{aligned} \tag{3}$$

In a heterogeneous treatment effect model, the β coefficients are to be understood as means or mean effects. For instance, $\beta_0 = E[Y(0,0)]$. Furthermore, note that $E[Y(0,1)] = \beta_0 + \beta_1$ cannot be identified in the data because D is multicollinear with its interaction with Z and would need to be dropped from an empirical regression. That is, when for instance comparing $Z = 0.25$ vs. $Z = 0$, β_1 cannot be separated from the interaction β_5. In contrast, differences in the interactions can in principle be identified over positive values z to investigate the effect heterogeneity in $\delta(z)$. Even if β_1 cannot be separated from β_5 (for $Z = 0.25$ vs. $Z = 0$) or β_6 (for $Z = 0.5$ vs. $Z = 0$), the difference of the empirical coefficients of $DI\{Z = 0.25\}$ and $DI\{Z = 0.5\}$ yield an estimate for $\beta_5 - \beta_6$ and thus, $\delta(0.25) - \delta(0.5)$. Likewise, $\theta(z, 0, 1) = E[Y(z,1) - Y(0,1)]$ for $z \in \{0.25, 0.5, 0.75\}$ is not identified, but $\theta(z', z, 1)$ for $z' \neq z$ and $z', z \in \{0.25, 0.5, 0.75\}$ is, as well as any $\theta(z', z, 0)$].

Summing up, the design of PROGRESA allows (without further functional form assumptions) to only consider effects in subgroups defined upon $D(z)$. Double randomization in Crépon, Duflo, Gurgand, Rathelot, and Zamora (2012) identifies the effects under particular z and d for the total population, however, for potentially infeasible combinations of the regional and individual treatments, unless the policy maker maintains the random assignment of D beyond

[10] However, further complications arise if not everyone assigned to a particular treatment state complies with the individual treatment assignment. In this case, populations defined upon individual treatment compliance may differ from the total population. Here, we do not consider such issues and focus on the (intention to treat) effect of treatment assignment.

the experiment.[11] For instance, $\delta(0.25)$ yields the average individual effect for the total population (i.e. for 100% of the individuals) when in fact only 25% are treated. A similar problem arises in Ferracci, Jolivet, and van den Berg (2010) who assume quasi-randomization of Z and D conditional on observed characteristics and also estimate $\delta(z)$ for various z. The approaches considered so far can therefore potentially be extended or improved upon by (i) considering further populations that are of policy interest and/or (ii) by basing the analysis on potential values of the individual treatment that would hypothetically occur for a particular regional treatment intensity.

2.3 Natural interference effects

To consider effects defined upon potential individual treatment states, we make use of insights from the literature on mediation analysis or natural direct and indirect effects in the denomination of Pearl (2001)).[12] Specifically, we define the interference effect conditional on the potential individual treatment state under a particular regional treatment intensity, denoted by $D(z)$, rather than setting it to a specific value d for every individual, which might be at odds with the regional treatment intensity apart from special cases. For a binary Z,

$$\theta_t(D(z)) = E[Y(1, D(z)) - Y(0, D(z))|T = t], \quad z \in \{0, 1\}. \tag{4}$$

For instance, $\theta_t(D(1))$ corresponds to the mean difference in potential outcomes when varying Z while fixing the individual treatment states in population t at their potential values for $z = 1$, rather than imposing the same value of D for everyone in population t.

Making use of the natural effects notation allows decomposing the total average treatment effect in the treated regions (ATET), which is defined as $\Delta_{Z=1} = E[Y(1, D(1)) - Y(0, D(0))|Z = 1]$ and is arguably of major policy interest. It comprises of both the effect operating through the individual treatment decision D, reflected by mean differences in outcomes induced by differences in $D(1)$ and $D(0)$ in treated regions, and the (natural) interference effect. We first consider the

[11] This appears unrealistic in many empirical contexts of policy interventions. Active labor market programs, for instance, are typically selective with respect to job seeker characteristics such as education and work experience.

[12] Robins and Greenland (1992) and Robins (2003) refer to such parameters as total or pure direct and indirect effects and Flores and Flores-Lagunes (2009) as net and mechanism average treatment effects.

latter and set $d = D(1)$:

$$\theta_{Z=1}(D(1)) \quad = \quad E[Y(1, D(1)) - Y(0, D(1))|Z = 1]. \tag{5}$$

This parameter corresponds to the average interference effect among the population in the treated regions (e.g. including both eligibles and ineligibles in villages treated by PROGRESA). Using the law of total probability, $\theta_{Z=1}(D(1))$ can be decomposed into the effects on various subgroups. In fact, it is a weighted average of interference effects on those for which $D(1) = 1$ and $D(1) = 0$ in treated regions:

$$\begin{aligned}
\theta_{Z=1}(D(1)) &= E[Y(1,1) - Y(0,1)|Z = 1, D(1) = 1] \cdot \Pr(D(1) = 1|Z = 1) \\
&+ E[Y(1,0) - Y(0,0)|Z = 1, D(1) = 0] \cdot \Pr(D(1) = 0|Z = 1) \\
&= \underbrace{E[Y(1,1) - Y(0,1)|Z = 1, D = 1]}_{\theta_{Z=1,D=1}(1)} \cdot \Pr(D(1) = 1|Z = 1) \\
&+ \underbrace{E[Y(1,0) - Y(0,0)|Z = 1, D = 0]}_{\theta_{Z=1,D=0}(0)} \cdot \Pr(D(1) = 0|Z = 1).
\end{aligned}$$

The second equality follows because $D(1) = D$ if $Z = 1$. $\theta_{Z=1,D=1}(1)$ and $\theta_{Z=1,D=0}(0)$, the interference effects for those taking and those not taking the treatment in the treated region, may differ across both groups and effect heterogeneity can come from two sources. First, it may be due to interaction effects between Z and D so that $\theta_{Z=1,D=d}(1) \neq \theta_{Z=1,D=d}(0)$ for $d \in \{1, 0\}$, as already mentioned before. Second, another form of effect heterogeneity arises if treated and non-treated individuals differ systematically in their outcome-relevant characteristics, implying that $\theta_{Z=1,D=1}(d) \neq \theta_{Z=1,D=0}(d)$. By noting that $\theta_{Z=1,D=1}(1) = \theta_{Z=1,\mathcal{T}\in\{c,a\}}(1)$ (because $D(1) = 1$ for c, a) and $\theta_{Z=1,D=0}(0) = \theta_{Z=1,\mathcal{T}\in\{d,n\}}(0)$ (because $D(1) = 0$ for d, n) such that the types differ in both parameters, effect heterogeneity appears plausible in many applications.[13] Only if the individual treatment is randomized so that both groups have similar characteristics or the interference effect is homogeneous (i.e. does not depend on individual characteristics), it generally follows that $\theta_{Z=1,D=1}(d) = \theta_{Z=1,D=0}(d)$. In the special case that both potential sources of effect

[13] In the case of PROGRESA where always takers and defiers are ruled out and Z is randomized, $\theta_{Z=1,D=1}(1) = \theta_c(1)$ and $\theta_{Z=1,D=0}(0) = \theta_n(0)$.

heterogeneity are nil, $\theta_{Z=1,D=1}(1) = \theta_{Z=1,D=0}(0) = \theta_{Z=1}(D(1)) = \theta_{Z=1}$.

Theoretically, we could also consider the interference effect when setting $d = D(0)$:[14]

$$\theta_{Z=1}(D(0)) \quad = \quad E[Y(1, D(0)) - Y(0, D(0))|Z = 1]. \tag{6}$$

However, this appears to be a less interesting parameter, because it assesses the interference effect among those with $Z = 1$ when hypothetically setting their individual treatments to the potential values that would occur under $Z = 0$.[15] Evaluation is then based on potential individual treatment values (among those with $Z = 1$) that are in general not consistent with the aggregate treatment intensity implied by $Z = 1$, a case that cannot occur in reality.[16]

2.4 Natural individual effects

Subtracting from the ATET the natural interference effect yields the part of the total treatment effect that operates through individual treatment assignment:

$$\begin{aligned} \Delta_{Z=1} - \theta_{Z=1}(D(1)) \quad &= \quad E[Y(1, D(1)) - Y(0, D(0))|Z = 1] - E[Y(1, D(1)) - Y(0, D(1))|Z = 1] \\ &= \quad E[Y(0, D(1)) - Y(0, D(0))|Z = 1] = \gamma_{Z=1}(0). \end{aligned} \tag{8}$$

This parameter, henceforth referred to as the natural individual effect, corresponds to the difference in mean potential outcomes in treated regions coming from hypothetically varying the potential individual treatment states from $D(1)$ to $D(0)$, if Z was set to zero. By fixing the regional treatment intensity at zero in the potential outcomes, any interference is switched off so that any impact is due to individual treatment effects (net of any interaction with Z). At first

[14] Vansteelandt and VanderWeele (2012) also discuss identification of (natural) effects on the population with $Z = 1$. Concerning $\theta_{Z=1}(d)$, they consider $d = D(1)$ rather than $d = D(0)$ as a natural reference among those with $Z = 1$ when the choice of reference levels appears a priori hard to justify.

[15] If Z=0 implies a zero intensity, then these potential values would be zero for everyone, i.e. D(0)=0, see the discussion above.

[16] For the same reason, assessing

$$\theta_{Z=1}(1) = E[Y(1, 1) - Y(0, 1)|Z = 1], \theta_{Z=1}(0) = E[Y(1, 0) - Y(0, 0)|Z = 1] \tag{7}$$

rather than $\theta_{Z=1,D=1}(1)$ and $\theta_{Z=1,D=0}(0)$ seems somewhat odd, because apart from the special case that for instance everyone is individually treated under $Z = 1$ ($\Pr(D(1) = 1|Z = 1) = 1$), the individual treatment distribution is inconsistent with the regional intensity implied by $Z = 1$.

glance, the natural individual effect may not be the most interesting parameter to policymakers, who might prefer to learn about the individual effect of a comparison of $D = 1$ vs. $D = 0$ rather than $D(1)$ vs. $D(0)$. The discussion nevertheless highlights that $\gamma_{Z=1}(0)$ and (the arguably more interesting) $\theta_{Z=1}(D(1))$ sum up to the total treatment effect in treated regions, $\Delta_{Z=1}$.

Furthermore, $\gamma_{Z=1}(0)$ may be disentangled into the contributions of the individual treatment effects of two compliance types:

$$\begin{aligned}\gamma_{Z=1}(0) &= \underbrace{E[Y(0,1) - Y(0,0)|Z=1, \mathcal{T}=c]}_{\delta_{Z=1,\mathcal{T}=c}(0)} \cdot \Pr(\mathcal{T}=c|Z=1) \\ &+ \underbrace{E[Y(0,0) - Y(0,1)|Z=1, \mathcal{T}=d]}_{\delta_{Z=1,\mathcal{T}=d}(0)} \cdot \Pr(\mathcal{T}=d|Z=1) \qquad (9)\end{aligned}$$

Only compliers and defiers contribute to $\gamma_{Z=1}(0)$, because $D(1) = D(0)$ for always and never takers such that their natural individual effect is zero by definition. Furthermore, if positive monotonicity of D in Z holds ($\Pr(D(1) \geq D(0)) = 1$, see Imbens and Angrist (1994)), then $\Pr(\mathcal{T} = d|Z = 1) = 0$, implying that (9) reduces to $\gamma_{Z=1}(0) = \delta_{Z=1,\mathcal{T}=c}(0) \cdot \Pr(\mathcal{T} = c|Z = 1)$. This appears reasonable in many empirical contexts and is naturally satisfied if the regional treatment intensity is exactly zero under $Z = 0$ such that $\Pr(D = 1|Z = 0) = 0$, as in PROGRESA. $\gamma_{Z=1}(0)$ resembles the so called intention-to-treat effect in policy evaluation. In fact, both parameters are the same in the absence of interference effects, i.e., if the individual level SUTVA holds.

Moreover, under positive monotonicity, $\delta_{Z=1,\mathcal{T}=c}(0) = \gamma_{Z=1}(0)/\Pr(Z = 1, \mathcal{T} = c)$, a result that resembles the literature on the local average treatment effect (LATE) on compliers: The natural individual effect scaled by the share of compliers gives the individual level treatment effect among the compliers, i.e., among those whose individual treatment state is actually affected by the regional treatment intensity, which appears to be an interesting target population. Note that under random assignment of Z, $\delta_{Z=1,\mathcal{T}=\tau}(0) = \delta_{\mathcal{T}=\tau}$ for $\tau \in \{c, d\}$. Under both monotonicity and random assignment, $\gamma_{Z=1}(0) = \delta_{\mathcal{T}=c}(0) \cdot \Pr(\mathcal{T} = c)$ and $\delta_{\mathcal{T}=c}(0) = \gamma_{Z=1}(0)/\Pr(\mathcal{T} = c)$.

In this context, one might wonder whether $\delta_{Z=1,\mathcal{T}=c}(1)$ is a more relevant parameter than $\delta_{Z=1,\mathcal{T}=c}(0)$, given that the former is defined upon the regional treatment that corresponds to the conditional one ($Z = 1$) while the latter is not. The answer is likely yes if one aims at assessing

individual level treatment effects and considers the regional treatment intensity (which can be regarded as confounder that needs to be controlled for) of $Z = 1$ as given/externally set. The answer is, however, likely no if the goal is to disentangle interference and individual level effects of some policy intervention. In this case, simultaneously considering $\theta_{Z=1}(D(1))$ and $\delta_{Z=1,\mathcal{T}=c}(1)$ would account twice for interactions between Z and D such that the parameters generally do not add up to the ATET. If one considers $\theta_{Z=1}(D(1))$ to be the most appropriate measure of interference effects (due to reasons outlined before), then $\delta_{Z=1,\mathcal{T}=c}(0)$ naturally follows as measure of the 'pure' (in the notation of Robins and Greenland (1992)) individual level treatment effect net of any impact of the regional treatment intensity. Specifically, if $Z = 0$ implies a regional treatment intensity of zero, $\delta_{Z=1,\mathcal{T}=c}(0)$ can be interpreted as the average individual treatment effect in the absence of any interference such that SUTVA holds on the individual level.

2.5 Further parameters

Theoretically, individual treatment effects could also be defined for populations different to those individually affected by Z, e.g. for the total population living in treated regions:

$$\delta_{Z=1}(z) = E[Y(z,1) - Y(z,0)|Z = 1]. \tag{10}$$

However, the regional intensity $Z = z$ does generally not correspond to an individual treatment prescription of $D = 1$ for everyone with $Z = 1$, implying that $\delta_{Z=1}(z)$ refers to a hypothetical setting that cannot be attained in reality (similarly to the evaluation of $\delta_{Z=1}$ in Ferracci, Jolivet, and van den Berg (2010)). Again, $\delta_{Z=1}(0)$ might nevertheless be interesting if one wants to infer on the individual level treatment effect when switching off interference, as it would be the case under individual-level SUTVA. Small-scale evaluations, i.e., randomized control trials, may come close to measuring this parameter. An intervention evaluated in a relatively small random sample of the population is unlikely to generate sizeable interference effects. Therefore, the sample average treatment effect (SATE), see Imbens (2004), in such evaluations might be an (almost) unbiased estimator of $\delta_{Z=1}(0)$. However, when an intervention is scaled up, interference effects might occur. In this case, the SATE is not an unbiased estimator of the average treatment effect

(ATE) in the total population.

Note that only in a very specific case, $\delta_{Z=1}(0)$ and the interference effect $\theta_{Z=1}(1)$ add up to the ATET, namely if $\Pr(\mathcal{T} = c|Z = 1) = 1$ such that everyone in the treated region is a complier (and thus, individually treated). It is easy to see from (9) that under this condition, $\gamma_{Z=1}(0) = \delta_{Z=1,\mathcal{T}=c}(0) = \delta_{Z=1}(0)$.

In line with previous arguments, $\delta_{Z=1}(1)$, $\delta_{Z=1}(0)$, $\delta_{Z=1,\mathcal{T}=\tau}(1)$, and $\delta_{Z=1,\mathcal{T}=\tau}(0)$ for $\tau \in \{c, d, a, n\}$ generally differ. Only if there are no interaction effects between Z and D, $\delta_t(1) = \delta_t(0)$ for any population t. If in addition, individual treatment effects do not vary with individual characteristics (i.e. are homogeneous), it holds that $\delta_t(1) = \delta_{Z=1}(0) = \delta_{Z=1,\mathcal{T}=\tau}(1) = \delta_{Z=1,\mathcal{T}=\tau}(0) = \delta_{Z=1}$ for $\tau \in \{c, d, a, n\}$. Under the no interactions assumption, effect homogeneity, and monotonicity of D in Z, it follows that

$$\gamma_{Z=1} \quad = \quad \delta_{Z=1} \cdot \Pr(\mathcal{T} = c|Z = 1).$$

In this particular case, the ATET can be decomposed into

$$\Delta_{Z=1} \quad = \quad \theta_{Z=1} + \gamma_{Z=1},$$

while in the general case without imposing these restrictions,[17] it can be seen from (8) that

$$\Delta_{Z=1} \quad = \quad \theta_{Z=1}(D(1)) + \gamma_{Z=1}(0) = \theta_{Z=1}(D(0)) + \gamma_{Z=1}(1). \tag{11}$$

Analogous to the ATET, also Δ, the ATE in the total population (comprising all subjects living both in treated and nontreated regions) may be decomposed into natural interference and individual effects $(\theta(D(1)), \gamma_{Z=1}(0), \theta_{Z=1}(D(0)), \gamma_{Z=1}(1))$, which is omitted for the sake of brevity. Note that $\Delta = \Delta_{Z=1}$ if Z is randomized, while more generally, $\Delta = \Delta_{Z=1} \cdot \Pr(Z = 1) + \Delta_{Z=0} \cdot \Pr(Z = 0)$. Whether it is useful to consider and decompose Δ in addition to $\Delta_{Z=1}$

[17] Note that even under constant $\theta_{Z=1}$ and $\gamma_{Z=1}$, the individual level SUTVA is generally violated: If $\theta_{Z=1}$ is non-zero, the overall effect of the treatment on the individual outcome depends on the share of subjects that receive the treatment and therefore, the treatment assignment to other subjects matters for $\Delta_{Z=1}$. Only if $\theta_{Z=1} = 0$ such that interference effects are ruled out, the SUTVA is satisfied at the individual level.

depends on the empirical context. If nontreated regions are never planned to be targeted by the policy intervention, the assessment of the ATE (which includes those with $Z = 0$) bears little relevance compared to the ATET. However, if the goal is to roll out the intervention to the entire population, decomposing Δ allows evaluating through which channels the intervention would affect the individual outcomes.

A further potentially interesting parameter is the average individual treatment effect on those individually treated in regions with high treatment intensity under a hypothetical $Z = 1$, in order to judge whether D itself was effective among those who received it:

$$\begin{aligned}
\delta_{Z=1,D=1}(1) &= E[Y(1,1) - Y(1,0)|Z=1, D=1] \\
&= \underbrace{E[Y(1,1) - Y(1,0)|Z=1, \mathcal{T}=c]}_{\delta_{Z=1,\mathcal{T}=c}(1)} \cdot \Pr(\mathcal{T}=c|Z=1, D=1) \\
&+ \underbrace{E[Y(z,1) - Y(z,0)|Z=1, \mathcal{T}=a]}_{\delta_{Z=1,\mathcal{T}=a}(1)} \cdot \Pr(\mathcal{T}=a|Z=1, D=1).
\end{aligned}$$

$\delta_{Z=1,D=1}(1)$ is a mixture of the impacts on the compliers and always takers, as for either group $D(1) = 1$, such that observing $Z = 1$ implies $D = 1$. If Z is randomly assigned, it follows that $\delta_{Z=1,\mathcal{T}=\tau}(1) = \delta_{\mathcal{T}=\tau}(1)$ and $\Pr(\mathcal{T} = \tau|Z = 1, D = 1) = \Pr(\mathcal{T} = \tau|D = 1)$ for $\tau \in \{c, a\}$. Moreover, if $\Pr(D = 1|Z = 0) = 0$ as in PROGRESA, always takers are ruled out and $\delta_{Z=1,D=1}(1) = \delta_{Z=1,\mathcal{T}=c}(1)$. Finally, if both the random assignment of Z and $\Pr(D = 1|Z = 0) = 0$ hold, $\delta_{Z=1,D=1}(1) = \delta_{\mathcal{T}=c}(1)$, the individual treatment effect on those individually treated in the high treatment intensity region corresponds to the individual treatment effect on the compliers in the population. In this case, the individual treatment and interference effects defined on the opposite state of Z sum up to the (local) average treatment effect on compliers (LATE),

$$\Delta_c = \gamma_c(1) + \theta_c(D(0)) = \delta_c(1) + \theta_c(0) = \delta_{Z=1,D=1}(1) + \theta_{Z=1,D=1}(0), \tag{12}$$

where $\gamma_c(1) = \delta_c(1)$ and $\theta_c(D(0)) = \theta_c(0)$ follows from the fact that $D(z) = z$ for the compliers.

If Z is defined such that $\Pr(D = 1|Z = 0) > 0$, also the individual treatment effect on those

individually treated in regions with low treatment intensity appear interesting:

$$\begin{aligned}\delta_{Z=0,D=1}(0) &= E[Y(z,1)-Y(z,0)|Z=0,D=1] \\ &= \underbrace{E[Y(0,1)-Y(0,0)|Z=0,\mathcal{T}=d]}_{\delta_{Z=0,\mathcal{T}=d}(0)}\cdot\Pr(\mathcal{T}=d|Z=0,D=1) \\ &+ \underbrace{E[Y(0,1)-Y(0,0)|Z=0,\mathcal{T}=a]}_{\delta_{Z=0,\mathcal{T}=a}(0)}\cdot\Pr(\mathcal{T}=a|Z=0,D=1).\end{aligned}$$

Under random assignment of Z, $\delta_{Z=0,\mathcal{T}=\tau}(0)=\delta_{\mathcal{T}=\tau}(0)$ and $\Pr(\mathcal{T}=\tau|Z=0,D=1)=\Pr(\mathcal{T}=\tau|D=1)$ for $\tau\in\{d,a\}$, under positive monotonicity of D in Z (see the discussion above), $\delta_{Z=0,D=1}(0)=\delta_{Z=0,\mathcal{T}=a}(0)$. Under both restrictions together, $\delta_{Z=0,D=1}(0)=\delta_{\mathcal{T}=a}(0)$. In contrast to $\delta_{Z=1,D=1}(1)$ and $\delta_{Z=0,D=1}(0)$, we argue that $\delta_{Z=1,D=1}(0)$ and $\delta_{Z=0,D=1}(1)$ are less interesting parameters, because they are defined on a practically infeasible combination of individual and regional treatment states. Note that the same argument generally applies to the parameters $\delta_{Z=1,D=0}(1)$ and $\delta_{Z=0,D=0}(0)$.

Table 1: Summary of effects

parameter	symbol	description
$E[Y(1,D(1))-Y(0,D(0))\|Z=1]$	$\Delta_{Z=1}$	(total) treatment effect in the treatment regions
$E[Y(1,D(z))-Y(0,D(z))\|Z=1]$	$\theta_{Z=1}(D(z))$	natural interference effect in the treatment regions
$E[Y(z,D(1))-Y(z,D(0))\|Z=1]$	$\gamma_{Z=1}(z)$	natural individual effect in the treatment regions
$E[Y(1,1)-Y(0,0)\|\mathcal{T}=c]$	Δ_c	(total) treatment effect on compliers
$E[Y(1,d)-Y(0,d)\|\mathcal{T}=c]$	$\theta_c(d)$	interference effect on compliers
$E[Y(z,1)-Y(z,0)\|\mathcal{T}=c]$	$\delta_c(z)$	individual treatment effect on compliers
$E[Y(1,d)-Y(0,d)\|\mathcal{T}=n]$	$\theta_n(d)$	interference effect on never takers
$E[Y(1,d)-Y(0,d)\|\mathcal{T}=a]$	$\theta_a(d)$	average interference effect on always takers
$E[Y(z,1)-Y(z,0)\|Z=z,D=1]$	$\delta_{Z=z,D=1}(z)$	ind. treatment effect cond. on individual treatment and $Z=z$
$E[Y(1,d)-Y(0,d)\|Z=z,D=d]$	$\theta_{Z=z,D=d}(d)$	interference effect cond. on $D=d$ and $Z=z$

Note: Note that $\theta_c(d)=\theta_c(D(z))$ and $\delta_c(z)=\gamma_c(z)$ if $d=z$ because $D(z)=z$ for $\mathcal{T}=c$.

Finally, policy makers might also want to learn about $\theta_{Z=1,D=1}(1)$ and $\theta_{Z=0,D=1}(1)$, i.e., know to which interference effects individually treated individuals in regions with high or low treatment intensity are exposed to under treatment.[18] Likewise, the interference effects on nontreated

[18]However, bear in mind that in the presence of always takers, (12) does not hold such that interference and individual treatment effects in the individually treated population do not add up to the total effect of the policy intervention.

individuals, $\theta_{Z=1,D=0}(0)$ and $\theta_{Z=0,D=0}(0)$, are also of interest allow learning about the effects of an intervention on individuals that are not individually treated. Table 1 summarizes the various effects we deem potentially interesting.

3 Identifying assumptions

In this section, we more formally discuss different sets of assumptions and their identifying power w.r.t. to the various effects of interest introduced in Section 2. Specifically, we consider the cases of (i) randomization of Z and deterministic assignment of D (as in PROGRESA), (ii) double randomization of Z and D (as in Crépon, Duflo, Gurgand, Rathelot, and Zamora (2012)), (iii) selection on observables w.r.t. Z and D (which is for instance related to Ferracci, Jolivet, and van den Berg (2010) and Frölich and Michaelowa (2011)), and (iv) identification based on difference-in-differences (as in Lalive, Landais, and Zweimüller (2015)).[19] Throughout the section, we maintain the following regional SUTVA, which rules out general equilibrium, spillover, and other interference effects across regions:

Assumption 1 (SUTVA on the regional level):
In any region k, any individual level potential outcome $Y(z,d)$ does not depend on the value of Z and the distribution of D in any other region $k \neq l$.

3.1 Randomization of Z and deterministic D

We subsequently formalize the assumptions of random regional treatment assignment and deterministic individual treatment assignment, assuming Z and D are binary.

Assumption 2 (random assignment of the regional treatment):
$\{Y(z',d), D(z)\} \perp Z$ for all $z', z, d \in \{0,1\}$.

[19] For instrumental variable strategies, see Frölich and Huber (2014) and references therein.

By Assumption 2,

$$\begin{aligned}\Delta_{Z=1} &= E[Y(1,D(1)) - Y(0,D(0))|Z=1] = E[Y(1,D(1)) - Y(0,D(0))] = \Delta \\ &= E[Y|Z=1] - E[Y|Z=0]. \end{aligned} \tag{13}$$

Assumption 3 (deterministic individual treatment assignment):

$D = g(Z, X)$, with g being a known function

In PROGRESA for instance, individual treatment is a deterministic function of Z and a poverty index, which is itself a deterministic function of observed characteristics X. Specifically, $D = g(0, X) = 0$, while $D = g(1, X)$ might be either 1 or 0 depending on the values in X that determine the score of the poverty index. From Assumption 3, which implies that D is a deterministic in Z, X, it also follows that $D(z)$ for $z \in \{1, 0\}$ and thus, the type $\mathcal{T}$ is identified. Combined with Assumption 2, it follows that interference effects on never takers and always takers (if they exist), as well the total effects on compliers and defiers are identified.[20]

$$\begin{aligned}\theta_n(0) &= E[Y(1,0) - Y(0,0)|\mathcal{T}=n] = E[Y(1,0)|Z=1,\mathcal{T}=n] - E[Y(0,0)|Z=0,\mathcal{T}=n] \\ &= E[Y|Z=1,\mathcal{T}=n] - E[Y|Z=0,\mathcal{T}=n], \end{aligned} \tag{14}$$

$$\begin{aligned}\theta_a(1) &= E[Y(1,1) - Y(0,1)|\mathcal{T}=a] = E[Y(1,1)|Z=1,\mathcal{T}=a] - E[Y(0,1)|Z=0,\mathcal{T}=n] \\ &= E[Y|Z=1,\mathcal{T}=a] - E[Y|Z=0,\mathcal{T}=a], \\ \Delta_c &= E[Y(1,1) - Y(0,0)|\mathcal{T}=c] = E[Y(1,1)|Z=1,\mathcal{T}=c] - E[Y(0,0)|Z=0,\mathcal{T}=c] \\ &= E[Y|Z=1,\mathcal{T}=c] - E[Y|Z=0,\mathcal{T}=c], \\ \Delta_d &= E[Y(1,1) - Y(0,0)|\mathcal{T}=d] = E[Y(1,1)|Z=0,\mathcal{T}=d] - E[Y(0,0)|Z=1,\mathcal{T}=d] \\ &= E[Y|Z=0,\mathcal{T}=d] - E[Y|Z=1,\mathcal{T}=d]. \end{aligned}$$

Disentangling the total effect on the compliers (and defiers) requires further assumptions. One potential approach consists of imposing effect homogeneity in the interference effects across types, then $\theta_c(0) = \theta_n(0)$, $\delta_c(1) = \Delta_c - \theta_n(0)$, $\theta_c(1) = \theta_a(1)$, $\delta_c(0) = \Delta_c - \theta_a(1)$. We refer to Forastiere, Mealli, and VanderWeele (2016) for a more thorough discussion of homogeneity assumptions on

[20] We acknowledge that the case of defiers is likely practically irrelevant.

potential outcomes or effects across types for identifying interference effects.

3.2 Randomization of the regional and individual treatment

We subsequently maintain Assumption 2, but replace Assumption 3 by random individual treatment assignment conditional on Z:

Assumption 4 (random assignment of the individual treatment within regions): $Y(z',d) \perp D|Z=z$, for all $z',z,d \in \{0,1\}$.

Under Assumptions 2 and 4, the individual effect among the individually treated is given by

$$\begin{aligned}\delta_{Z=z,D=1}(z) &= E[Y(z,1)-Y(z,0)|Z=z,D=1] = E[Y(z,1)-Y(z,0)] = \delta(z) \\ &= E[Y|Z=z,D=1]-E[Y|Z=z,D=0] \\ &= \frac{E[Y\cdot D|Z=z]}{\Pr(D=1|Z=z)} - \frac{E[Y\cdot(1-D)|Z=z]}{1-\Pr(D=1|Z=z)},\end{aligned} \tag{15}$$

where the last equality follows from basic probability theory. Therefore, this parameter is identified whenever the following overlap condition is satisfied:

Assumption 5 (existence of individuals with $D=1$ and $D=0$ conditional on $Z=z$): $0 < \Pr(D=1|Z=z) < 1$.

As already discussed in Section 2, common support is for instance violated in PROGRESA for $z=0$, where nobody is individually treated in non-treatment regions: $\Pr(D=1|Z=0)=0$. In such a case, $\delta(0)$ cannot be nonparametrically identified.

Furthermore, under Assumptions 2 and 4, the interference effect conditional on Z,D is

$$\begin{aligned}\theta_{Z=z,D=d}(d) &= E[Y(1,d)-Y(0,d)|Z=z,D=d] = E[Y(1,d)-Y(0,d)] = \theta(d) \\ &= E[Y|Z=1,D=d]-E[Y|Z=0,D=d] \\ &= \frac{E[Y\cdot Z|D=d]}{\Pr(Z=1|D=d)} - \frac{E[Y\cdot(1-Z)|D=d]}{1-\Pr(Z=1|D=d)},\end{aligned} \tag{16}$$

conditional on the satisfaction of the following overlap condition:

Assumption 6 (existence of regions with $Z = 1$ and $Z = 0$ conditional on $D = d$):
$0 < \Pr(Z = 1|D = d) < 1$.

Assumption 6 is of course closely linked to Assumption 5 (albeit tailored to $\theta(d)$ rather than $\delta(z)$). By Bayes' theorem, $\Pr(D = 1|Z = 0) = 0$ for instance implies that $\Pr(Z = 0|D = 1) = 0$ such that $\Pr(Z = 1|D = 1) = 1$ and both Assumptions 5 and 6 are violated. $\theta(1)$ is therefore not identified in PROGRESA. Note that for the natural interference effect, $\theta(D(z))$, Assumption 6 becomes $0 < \Pr(Z = 1|D = D(z)) < 1$. If $D(z) = 1$ for some individuals and $D(z) = 0$ for others, common support needs to hold for both $D = 1$ and $D = 0$:

Assumption 7 (existence of regions with $Z = 1$ and $Z = 0$ given $D = 1$ and $D = 0$):
$0 < \Pr(Z = 1|D = d) < 1$ for all $d \in \{1, 0\}$.

It is easy to see from Bayes' theorem that Assumption 7 implies both Assumptions 5 and 6.

Under Assumptions 2, 4, and 7, the natural interference effect in the treatment regions is obtained by

$$\theta_{Z=1}(D(z)) = E\left[\left(\frac{Y \cdot Z}{\Pr(Z = 1|D)} - \frac{Y \cdot (1 - Z)}{1 - \Pr(Z = 1|D)}\right) \cdot \frac{\Pr(Z = z|D)}{\Pr(Z = z)}\right]. \tag{17}$$

Proof: See Appendix A.1.

Note that due to the random assignment of Z and D, $\theta_{Z=1}(D(z)) = \theta(D(z))$. Likewise, $\Delta_{Z=1} = \Delta$, $\theta_{Z=z,D=d}(d) = \theta(d)$, $\delta_{Z=z,D=d}(z) = \delta(z)$, and $\gamma_{Z=1}(z) = \gamma(z)$. Also the latter parameter is identified under Assumptions 2, 4, and 7, because by (11), $\gamma_{Z=1}(z) = \Delta_{Z=1} - \theta_{Z=1}(D(1 - z))$, which can be shown to correspond to

$$\gamma(z) \quad = \quad E\left[\frac{Y \cdot I\{Z = z\}}{\Pr(Z = z|D)} \cdot \left(\frac{\Pr(Z = 1|D)}{\Pr(Z = 1)} - \frac{1 - \Pr(Z = 1|D)}{1 - \Pr(Z = 1)}\right)\right]. \tag{18}$$

This follows from the fact that (18) and (17) defined upon opposite values of z add up to $\Delta_{Z=1}$, which is $E\left[\frac{Y \cdot Z}{\Pr(Z=1)} - \frac{Y \cdot (1-Z)}{1-\Pr(Z=1)}\right] = E[Y|Z = 1] - E[Y|Z = 0]$ under Assumption 2. As a final remark, note that in a regression discontinuity design (see Hahn, Todd, and van der Klaauw (2001)), in which the assignment of D is non-random but discontinuously changes at a threshold of some index while regional treatment Z is randomized, the aforementioned assumptions may

approximately hold for a local population around the index threshold. In this case, the parameters of interest are identified for this local population, which suggests yet another evaluation approach to research designs similar to PROGRESA.[21]

3.3 Selection on observables

As a relaxation of the assumptions in Section 3.2, we now assume that regional treatment assignment is quasi-random conditional on a set of observables, denoted by X, and that individual treatment assignment is quasi-random conditional on Z, X. Such or similar sequential exogeneity assumptions have been considered by Pearl (2001), Flores and Flores-Lagunes (2009), Imai, Keele, and Yamamoto (2010), and Huber (2014), among many others. We therefore replace Assumptions 2 and 4 by Assumptions 8 and 9.

Assumption 8 (conditional independence of the regional treatment):

$\{Y(z', d), D(z)\} \perp Z | X = x$ for all $z', z, d \in \{0, 1\}$ and x in the support of X.

Assumption 8 states that the joint distribution of the potential outcomes and individual treatments are independent of the regional treatment intensity conditional on X. This rules out unobserved confounders affecting regional treatment assignment on the one hand and the potential outcomes and/or individual treatment under $z = 0$ on the other hand, when controlling for X.[22]

Assumption 9 (conditional independence of the individual treatment):

$Y(z', d) \perp D | Z = z, X = x$ for all $z', z, d \in \{0, 1\}$ and x in the support of X.

Assumption 9 states that the individual treatment is conditionally independent of the potential outcomes conditional on the actual regional treatment intensity Z and covariates X. This rules out unobserved confounders jointly affecting the individual treatment and the potential outcomes under $z = 0$ after controlling for X and Z.

Assumption 10 (common support restrictions):

[21] Angelucci and Di Maro (2015) consider a different regression discontinuity framework where Z rather than D is discontinuous at the threshold of some index.

[22] This is known as conditional independence, selection on observables, or exogeneity in the treatment evaluation literature, see for instance Imbens (2004).

(a) $\Pr(Z=1|X=x)<1$ for all x in the support of X,

(b) $0<\Pr(D=1|Z=z,X=x)<1$ for all x in the support of X,

(c) $0<\Pr(Z=1|D=d,X=x)<1$ for all x in the support of X,

(d) $0<\Pr(Z=1|D=d,X=x)<1$ for all $d\in\{1,0\}$ and x in the support of X.

Assumption 10(a) is the conventional common support assumption for the identification of the ATET, implying that there must not be any combination of covariates X that entails the higher regional treatment intensity with probability one (in order to find for any observation with $Z=1$ suitable matches with $Z=0$ that are comparable in X). It is satisfied under unconditional or stratified randomization of regional treatment intensity. Assumptions 10(b), 10(c), and 10(d) are analogous to Assumptions 5, 6, and 7, but are stronger in the sense that they require that the respective assumption holds conditional on all possible values of X. Similarly as before, Assumption 10(d) is stronger than and thus implies Assumptions 10(a), (b), and (c). Note that Assumption 10(b) is necessarily violated if individual treatment assignment depends deterministically on (an index of) X, as it is the case in PROGRESA.

Under Assumptions 8, 9, and 10(a),

$$\Delta_{Z=1}=E\left[\frac{Y\cdot Z}{\Pr(Z=1)}-\frac{Y\cdot(1-Z)}{1-\Pr(Z=1|X)}\cdot\frac{\Pr(Z=1|X)}{\Pr(Z=1)}\right], \tag{19}$$

see Hirano, Imbens, and Ridder (2003). Under Assumptions 8, 9, and 10(b),

$$\delta_{Z=z,D=1}(z)=E\left[\left(\frac{Y\cdot D}{\Pr(D=1|Z=z,X)}-\frac{Y\cdot(1-D)}{1-\Pr(D=1|Z=z,X)}\right)\cdot\frac{\Pr(Z=z|X)\cdot\Pr(D=1|Z=z,X)}{\Pr(Z=z)\cdot\Pr(D=1|Z=z)}\right], \tag{20}$$

while under Assumptions 8, 9, and 10(c),

$$\theta_{Z=z,D=d}(d)=E\left[\left(\frac{Y\cdot Z}{\Pr(Z=1|D=d,X)}-\frac{Y\cdot(1-Z)}{1-\Pr(Z=1|D=d,X)}\right)\cdot\frac{\Pr(Z=z|X)\cdot\Pr(D=d|Z=z,X)}{\Pr(D=d)\cdot\Pr(Z=z|D=d)}\right], \tag{21}$$

and finally, under Assumptions 8, 9, and 10(d),

$$\theta_{Z=1}(D(z))=E\left[\left(\frac{Y\cdot Z}{\Pr(Z=1|D,X)}-\frac{Y\cdot(1-Z)}{1-\Pr(Z=1|D,X)}\right)\cdot\frac{\Pr(Z=z|D,X)}{\Pr(Z=z|X)}\cdot\frac{\Pr(Z=1|X)}{\Pr(Z=1)}\right]. \tag{22}$$

The proofs for (20), (21), and (22) are provided in Appendix A.2 and are closely related to those

in Huber (2014). Furthermore, by subtracting $\theta_{Z=1}(D(1-z))$ from $\Delta_{Z=1}$, it can be shown that

$$\gamma_{Z=1}(z) = E\left[\frac{Y \cdot I\{Z=z\}}{\Pr(Z=z|D,X)} \cdot \left(\frac{\Pr(Z=1|D,X)}{\Pr(Z=1|X)} - \frac{1-\Pr(Z=1|D,X)}{1-\Pr(Z=1|X)}\right) \cdot \frac{\Pr(Z=1|X)}{\Pr(Z=1)}\right]. \tag{23}$$

Under particular conditions, one can even identify type-specific effects (Δ_c, $\theta_c(d)$, $\delta_c(z)$, $\theta_n(d)$, $\theta_a(d)$) despite the fact that in the absence of Assumption 3, Assumptions 8 and 9 do not permit learning the types from the data. Our assumptions imply that heterogeneity in potential outcomes across types is exclusively driven by X or more formally, that $Y(z',d) \perp D(z)|X=x$, see the discussion in Imai, Keele, and Yamamoto (2010), such that types and potential outcomes are conditionally independent given X. Therefore, appropriately averaging over X yields the effects on compliers, always takers, and never takers, under the additional assumption that D is monotonic in Z given X, which allows identifying the type proportions, see Abadie (2003):

Assumption 11 (monotonicity):

$\Pr(D(1) \geq D(0)|X=x) = 1$ for all x in the support of X.

In analogy to the concept of weighted treatment effects in Hirano, Imbens, and Ridder (2003), reweighing observations according to the distribution on X for the target population yields the parameter of interest. Consider, for instance, the weighted interference effect,

$$\theta_\omega(d) = E\left[\left(\frac{Y \cdot Z}{\Pr(Z=1|D=d,X)} - \frac{Y \cdot (1-Z)}{1-\Pr(Z=1|D=d,X)}\right) \cdot \frac{\omega(X)}{E[\omega(X)]}\right], \tag{24}$$

where $\omega(X)$ is the weighting function. By setting, for instance, $\omega(X) = 1 - \frac{D(1-Z)}{\Pr(Z=0|X)} - \frac{(1-D)Z}{\Pr(Z=1|X)}$, $\theta_c(d)$ is identified, as this approach reweighs observations according to the distribution of X among compliers, see Abadie (2003). The corresponding weighting functions for always and never takers are $\frac{D(1-Z)}{\Pr(Z=0|X)}$ and $\frac{(1-D)Z}{\Pr(Z=1|X)}$, respectively. Of course, populations other than the types could be considered, too, which would not require invoking Assumption 11. It is for instance easy to see that setting $\omega(X) = \Pr(Z=z, D=d|X)$ in (24) yields the result in (21).[23]

[23]The proof of (24) is omitted due to this analogy.

3.4 Difference-in-Differences

We now assume that the outcome variable is observed both in a baseline period prior to the assignment of Z and D and in the follow-up period after the assignment of Z and D when the effects are evaluated. Therefore, we introduce a time index for the period in which the outcome is measured: Y_0 denotes the observed pre-treatment outcome, while Y_1 is the outcome in the follow-up period. Note that in the previous discussion, $Y_1 = Y$. In contrast to randomization or quasi-randomization given observables, identification will subsequently rely on combining common trend assumptions on outcome changes over time with deterministic individual treatment assignment (Assumption 3) to identify effects on specific types. As in the standard difference-in-differences (DiD) framework, the identification results can be applied both to panel data and repeated cross sections.

Assumption 12 (common trends within types across regions):
$E[Y_1(0,0) - Y_0(0,0)|Z = z, \mathcal{T} = \tau] = E[Y_1(0,0) - Y_0(0,0)|Z = z', \mathcal{T} = \tau]$ for all $z \neq z'$ and $\tau \in \{n, c\}$.

Assumption 12 states that the mean potential outcomes in the absence of any regional and individual treatment within a particular type would change by the same magnitude from the pre-treatment to the follow-up period across (actual) regional treatment intensities.

Assumption 13 (no anticipation effect):
$\Pr(Y_0(z,d) = Y_0(0,0)) = 1$ for any z, d in the support of D, Z.

Assumption 13 rules out any anticipation effects of Z or D on the outcome in the baseline period.

Assuming that Z and D are binary, it follows from Assumptions 3, 12, and 13, that the interference effect among never takers and the total effect among compliers in the treated regions are identified:

$$\theta_{n,Z=1}(0) = E[Y_1|Z=1,\mathcal{T}=n] - E[Y_0|Z=1,\mathcal{T}=n] - [E[Y_1|Z=0,\mathcal{T}=n] - E[Y_0|Z=0,\mathcal{T}=n]], \quad (25)$$

$$\Delta_{c,Z=1} = E[Y_1|Z=1,\mathcal{T}=c] - E[Y_0|Z=1,\mathcal{T}=c] - [E[Y_1|Z=0,\mathcal{T}=c] - E[Y_0|Z=0,\mathcal{T}=c]].$$

Proof: See Appendix A.3.

In our application presented in Section 4, we will apply this strategy to reinvestigate labor market data from Lalive, Landais, and Zweimüller (2015). Finally, we present an assumption that allows evaluating interference effects among compliers under the condition that there are multiple regional treatment intensities. To keep the notation simple, suppose that Z can take three values: 0 (no regional treatment), 1 (low regional treatment intensity), and 2 (high regional treatment intensity). Furthermore, we state Assumption 3 more precisely by specifying that $g(z', X) = g(z, X)$ for any $z', z > 0$. This implies that eligibility depends on the same criteria in regions with high and low treatment intensity. Therefore, compliers satisfy $D(0) = 0, D(1) = D(2) = 1$ in this framework.

Assumption 14 (effect homogeneity among compliers across regions):

$E[Y_1(1,1) - Y_1(0,0)|Z = 2, \mathcal{T} = c] = E[Y_1(1,1) - Y_1(0,0)|Z = 1, \mathcal{T} = c]$.

Assumption 14 states that the average total effect among compliers of low regional treatment intensity vs. no regional treatment is homogeneous across high and low treatment regions.[24] This allows separating the total effect of $Z = 2$ vs. $Z = 0$ among compliers in high treatment regions into the total effect of $Z = 1$ vs. $Z = 0$, which by Assumption 14 is equal to the respective effect in low treatment regions, and the interference effect of $Z = 2$ vs. $Z = 1$. Under Assumptions 3, 12, 13, and 14, the spillover effect on compliers is identified by

$$\begin{aligned} \theta_{c,Z=2}(z' = 2, z = 1, d = 1) &= E[Y_1|Z = 2, \mathcal{T} = c] - E[Y_0|Z = 2, \mathcal{T} = c] \\ &- [E[Y_1|Z = 1, \mathcal{T} = c] - E[Y_0|Z = 1, \mathcal{T} = c]]. \end{aligned} \tag{26}$$

Proof: See Appendix A.3.

To increase their plausibility, the DiD assumptions might be relaxed to only hold conditional on observed covariates W. In this case, the effects are obtained by performing the DiD approach conditional on the covariates and averaging over the latter in a way that mimics the covariate distribution in the follow-up period of the respective type in treated regions. For instance, under

[24] Note, however, that the levels of the mean potential outcomes remain unrestricted.

the conditional satisfaction of Assumptions 3, 12, and 13 given W, the expressions in (25) become:

$$\begin{aligned}
\theta_{n,Z=1}(0) &= E_{W_1|Z=1,\mathcal{T}=n}\left[E[Y_1|W_1, Z=1, \mathcal{T}=n] - E[Y_0|W_0, Z=1, \mathcal{T}=n]\right. \\
&- \left.[E[Y_1|W_1, Z=0, \mathcal{T}=n] - E[Y_0|W_0, Z=0, \mathcal{T}=n]]\right], \qquad (27) \\
\Delta_{c,Z=1} &= E_{W_1|Z=1,\mathcal{T}=c}\left[E[Y_1|W_1, Z=1, \mathcal{T}=c] - E[Y_0|W_0, Z=1, \mathcal{T}=c]\right. \\
&- \left.[E[Y_1|W_1, Z=0, \mathcal{T}=c] - E[Y_0|W_0, Z=0.\mathcal{T}=c]]\right].
\end{aligned}$$

W_1, W_0 denote the observed covariates in periods 0 and 1, respectively. One possible identification approach is reweighting type-specific groups with observed $(W_0, Z=1)$, $(W_1, Z=0)$, and $(W_0, Z=0)$ towards the target group with $(W_1, Z=1)$ based on inverse probability weighting in a similar way as outlined in Abadie (2005). This requires the common support assumption that covariate values appearing in the target group with $(W_1, Z=1)$ also occur in the remaining three groups.

Finally, we refer to Deuchert, Huber, and Schelker (2016) for alternative DiD approaches to the identification of type-specific effects when Z is (in contrast to the present framework) randomized, but the compliance type is not directly observed, i.e., not a deterministic function of Z and X as imposed by Assumption 3.

4 Application

Lalive, Landais, and Zweimüller (2015) – henceforth LLZ – study market externalities of a large-scale extension of unemployment benefits and find that this policy decreased the job-search duration of ineligible individuals. We use the same setting and build on their work to illustrate the proposed framework in an empirical application. Our framework allows a refined interpretation of the effects in LLZ. Furthermore, we estimate the effects under a weaker common trend assumption than originally considered in several of their estimations. We also discuss testable implications of the stronger common trend assumption.

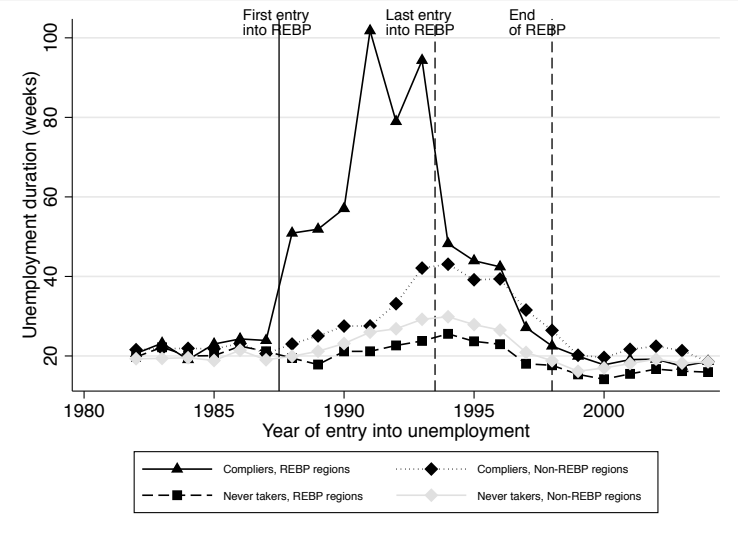
LLZ define as the *micro effect* changes in the search strategy of unemployed workers induced

by changes in unemployment insurance generosity. They refer to *market externalities* as changes in equilibrium labor market conditions induced by changes in unemployment insurance policies. The authors exploit a quasi-experimental setting to identify the effect of an UI benefit variation in a given labor market. They study job search outcomes of ineligible workers but who search in the same labor market. The aggregate unit, in this context the relevant labor market, is defined as the place where workers are competing for the same vacancies and is empirically determined. The Regional Extension Benefit Program (REBP) in Austria extended unemployment benefits drastically (an extra of three years) for a large subset of workers in selected regions of Austria from June 1988 until August 1993. This large UI benefits extension generated a strong increase in unemployment duration of treated workers thereby manipulating equilibrium labor market conditions. LLZ focus on unemployed workers in REBP regions who compete for the same vacancies and are similar to eligible workers, but are not eligible for REBP because they fail to meet the eligibility requirements of the REBP program. Using a difference-in-differences identification strategy, they compare these ineligible unemployed to unemployed in non-REBP regions to identify the effect of REBP on duration of job search of ineligible unemployed in treated markets.

Individual eligibility is defined as being above 50 at the start of their spell and having more than 15 years of work history in the past 25 years. Ineligible unemployed are those who were below 50 at the start of their spell or who have worked less than 15 out of the previous 25 years. In line with our conceptual framework, we refer to individuals fulfilling the individual criteria as *compliers* ($\mathcal{T} = c$) and to those not fulfilling the criteria as *never takers* ($\mathcal{T} = n$). Regional treatment (Z) is 1 for counties covered by REBP and 0 otherwise. The individual treatment D is the actual availability of the extended UI benefits to a particular subject. D equals one if $(Z = 1, \mathcal{T} = c)$ and zero otherwise. The definition of individual eligibility allows us to identify individuals in non-treated regions who would receive the treatment if they lived in a treated region. Figure 1 shows the average unemployment duration of compliers and never takers by residence in REBP and non-REBP counties by year of entry into unemployment. Unemployment duration was relatively stable and similar for all four groups between 1981 and 1986. In 1987 when REBP was enacted, unemployment duration rose sharply for compliers in REBP counties. The

difference to compliers in non-REBP counties can be interpreted as the total effect of the program on compliers. The other notable feature of the graph is the lower unemployment duration of never takers in REBP counties relative to never takers in non-REBP counties. Under our identifying assumptions discussed below, this difference represents the interference effect on never takers.

Figure 1: Unemployment duration by eligibility and REBP status of the county by year of entry into unemployment



Notes: The figure depicts the average unemployment duration in weeks for four distinct groups by year of entering unemployment. The first group are individuals in REBP counties who fulfill the REBP eligibility criteria of being above 50 and having more than 15 years of work history in the past 25 years prior to becoming unemployed. The second group are individuals who would fulfill the eligibility criteria but do not live in REBP counties. The third group are individuals in REBP counties who do not fulfill the eligibility criteria (less than 50 and/or less than 15 years of continuous work history). The fourth group are individuals who do not live in REBP counties and do not fulfill the eligibility criteria. We refer to individuals fulfilling the individual criteria as *compliers* and to those not fulfilling the criteria as *never takers*. The sample includes all men between 46 and 54 years who became unemployed in a given year. Non-REBP counties with high labor market integration to REBP counties are excluded from the sample.

In the notation of the current paper, LLZ identify the total effect on compliers ($\Delta_{c,Z=1}$) and the interference effect on never takers ($\theta_{n,Z=1}(0)$) in treated regions. We also aim for these parameters based on the identification result presented in equation (25), which requires the satisfaction of Assumptions 3, 13, and 12, i.e. common trends within types across regions. While this allows for different trends between types, several estimations in LLZ are based on the stronger common trend assumption ($E[Y_1(0,0) - Y_0(0,0)|Z = z] = E[Y_1(0,0) - Y_0(0,0)|Z = z']$) imposing common trends across types.[25] The latter implies the use all individuals in non-REBP counties as counterfactuals

[25]In most cases, LLZ assume this assumption to hold conditional on a set of observed characteristics, see their equation (2). In addition, however, LLZ also consider an approach that is in line with our Assumption 12, see their

for both compliers and never takers in treated regions.[26] This assumption has the testable implication that the difference in unemployment duration of compliers and never takers in non-REBP counties should remain constant over time, given that the assumptions underlying (25) are satisfied.

Table 2: Total effect on compliers and interference effect on never takers

	$\Delta_{c,Z=1}$		$\theta_{n,Z=1}(0)$		Test of A12'
	A12'	A12	A12'	A12	
Panel 1: by year of entering unemployment					
1988	25.67**	24.60**	-3.03	-2.42	1.69
	(11.93)	(12.01)	(2.95)	(2.84)	(1.27)
1989	25.07***	23.53***	-6.25	-5.08	2.71
	(6.47)	(6.93)	(4.27)	(4.02)	(1.65)
1990	27.92***	26.25***	-5.27	-3.73	3.21**
	(6.70)	(7.11)	(4.73)	(4.45)	(1.59)
1991	72.02***	70.94***	-5.88	-4.48	2.47
	(12.00)	(12.30)	(5.07)	(4.73)	(1.58)
1992	44.84***	42.52***	-8.75*	-5.96	5.11***
	(8.24)	(8.53)	(4.74)	(4.46)	(1.69)
1993	55.78***	48.93***	-12.07**	-4.82	14.09***
	(8.67)	(9.35)	(4.94)	(4.74)	(4.13)
Panel 2: average effect 1988-1993					
Average effect	50.38***	47.31***	-7.15**	-4.31	5.92***
1988 - 1993	(6.52)	(6.79)	(3.41)	(3.24)	(1.59)
Panel 3: average effect 1988-1993 - conditional on observables					
Average effect	48.54***	44.83***	-7.50***	-4.46	6.81***
1988 - 1993	(6.29)	(6.52)	(3.26)	(3.13)	(1.73)
Observations	59282	33581	56716	35403	47014

Notes: All panels present estimates of $\Delta_{c,Z=1}$ (columns 1 and 2) and $\theta_{n,Z=1(0)}$ (columns 3 and 4) and a test of the stronger common trend assumption across types (column 5). Results are presented for the main REBP period 1988-1993. The pre-treatment reference year is 1987. Panel 1 presents nonparametric estimates separately by year of entering unemployment. Panel 2 presents nonparametric estimates of the average effects over the period 1988-1993. Panel 3 presents semi-parametric estimates of the average effects over the period 1988-1993 conditional on education, family status, and tenure. A12 refers to the common trend assumption within types (Assumption 12) introduced in Section 3.4. A12' refers to the stronger common trend assumption ($E[Y_1(0,0) - Y_0(0,0)|Z = z] = E[Y_1(0,0) - Y_0(0,0)|Z = z']$). Standard errors clustered at the year x region level in parentheses. Standard errors in Panel 3 stem from a clustered bootstrap with 499 replications. *** denotes statistical significance at the 1 percent level, ** at the 5 percent level, and * at the 10 percent level. The estimation samples includes male workers between 46 and 54 years that were not employed in the steel sector. All duration outcomes are expressed in weeks.

equation (3).

[26] In this specific context, Assumption 3 could be relaxed to only hold in treated regions as the research design does not require distinguishing between compliers and never takers in the non-treated regions.

Table 2 presents the results when either invoking Assumptions 3 and 13 and a) the stronger common trend assumption (A12') requiring that $(E[Y_1(0,0) - Y_0(0,0)|Z = z] = E[Y_1(0,0) - Y_0(0,0)|Z = z'])$ or b) Assumption 12 (A12), i.e. common trends within types, in our evaluation sample. The latter differs from the original data in LLZ in several dimensions. First, we only investigate the period 1988-1993, whereas LLZ in addition consisder 1994-1997 when no new entries into the program occurred, but market externalities still persisted. Second, we only use 1987 as the untreated reference year, whereas LLZ exploit both years before and after the program as untreated reference years. Finally, we do not control for covariates whereas LLZ include covariates in their estimations.

Panel 1 of Table 2 provides the estimates by year of entering unemployment. The first two columns show the total effect on compliers ($\Delta_{c,Z=1}$). Columns 3 and 4 show the interference effect on never takers ($\theta_{n,Z=1}(0)$). The total effect of REBP on unemployment duration of compliers is strongly positive, i.e., eligible workers in treated regions substantially reduce their labor supply. The interference effect on unemployment duration of never takers is negative, pointing to the existence of a market externality. For both effects, the estimates are larger in absolute terms when A12' rather than A12 is invoked. The last column presents the difference in unemployment duration between compliers and never takers in non-REBP counties. Under A12', no differences between the two groups should exist. However, we see that the unemployment duration of compliers increases more over time than the unemployment duration of never takers. The stronger common trend assumption is thus unlikely to hold and the results under the common trend assumption within types appear more credible.

Panel 2 of Table 2 presents averages of these effects for the period 1988-1993. The average total effect on compliers is 50.38 weeks under A12' and about three weeks less under A12. The average interference effect on never takers is -7.15 under A12' and -4.31 weeks under A12. The latter is not statistically significant at conventional levels but of the same magnitude as the effect reported in LLZ under a somewhat different evaluation approach. Panel 3 presents estimates for the same effects as Panel 2, however, based on a conditional DiD approach as outlined in (3.4) by controlling for the covariates education, family status, and tenure. Applying semi-parametric

inverse probability weighting,[27] see the discussion at the end of Section 3.4, observations are reweighted to match the covariate distribution of the target population, i.e., compliers in columns 1 and 2 and never takers in columns 3 and 4, in REBP counties in the treatment period. In column 5, the target population are unemployed that meet the individual eligibility criteria in non-REBP counties in the treatment period. All results are very similar to Panel 2.

5 Conclusion

Most contributions in the field of treatment evaluation rule out general equilibrium, spillover, or interaction effects related to individual treatment assignment. This can be formalized by the Stable Unit Treatment Value Assumption (SUTVA), which assumes away any form of treatment-dependent interference between study participants, which, however, likely occurs in many empirical problems as for instance the assessment of labor market, development, or educational interventions. For this reason, this paper suggests a general framework for disentangling individual level treatment effects and interference effects under the crucial assumptions that SUTVA holds on an aggregate rather than individual level, e.g. across schools or regions. Borrowing notation from the causal mediation literature, we define a range of policy-relevant effects and formally discuss identification based on randomization, selection on observables, and difference-in-differences. Our approach therefore appears useful in various strands of applied research. In observational studies, it provides strategies for the identification of individual and interference effects. Furthermore, it may guide the design of experimental studies that aim on the one hand to disentangle individual and interference effects and on the other hand to investigate how the results of a small-scale randomized experiment develop when an intervention is scaled-up to a larger population.

As an empirical illustration, we reconsider data from Lalive, Landais, and Zweimüller (2015) who study the spillover effects of a large-scale extension of unemployment benefits in selected regions of Austria and find that this policy decreased the job-search duration of ineligible individuals in treated regions. Our framework provides a sharper definition of the identified effects. Furthermore, we apply our difference-in-differences methodology to identify the total effects on

[27] Conditional probabilities in our semi-parametric weighting approach are based on probit specifications.

eligibles and spillover effects on ineligibles under a somewhat weaker form of common trend assumption than underlying some of the results in Lalive, Landais, and Zweimüller (2015). Even though the estimates under either common trend assumptions are statistically significantly different, they qualitatively point into the same direction of a strong positive effect on the job-search duration of eligibles and a negative spillover effect among ineligibles.

A non-trivial question beyond the scope of this paper is how the boundaries of aggregate units should be defined. Regional SUTVA is only plausible if the aggregate units coincide with the relevant markets. Pre-defined regional or administrative entitites or cells formed for example by industry, education, or age as used in most empirical studies might only crudely approximate relevant markets. In a recent contribution, Nimczik (2016) provides a data-driven method to define labor markets and shows that traditional definitions perform quite poorly in separating distinct labor markets. Future research should therefore make use of the increasing availability of micro-data and advances in econometric modeling to implement and enhance data-driven approaches for the definition of relevant markets.

References

Abadie, A. (2003): "Semiparametric instrumental Variable estimation of treatment response models," *Journal of Econometrics*, 113, 231–263.

Abadie, A. (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72(1), 1–19.

Angelucci, M., and V. Di Maro (2015): "Program evaluation and spillover effects," *working paper, University of Michigan*.

Angelucci, M., and G. D. Giorgi (2009): "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?," *American Economic Review*, 99, 486–508.

Angelucci, M., S. Prina, H. Royer, and A. Samek (2015): "When incentives backfire: Spillover effects in food choice," *working paper, University of Michigan*.

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

Baird, S., A. Bohren, C. McIntosh, and B. Ozler (2012): "Designing Experiments to Measure Spillover and Treshold Effects," *discussion paper*.

Bobonis, G. J., and F. Finan (2009): "Neighborhood Peer Effects in Secondary School Enrollment Decisions," *Review of Economics and Statistics*, 91, 695–716.

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2012): "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment," Working Paper 18597, National Bureau of Economic Research.

Dahl, G. B., K. V. Løken, and M. Mogstad (2014): "Peer Effects in Program Participation," *American Economic Review*, 104, 2049–2074.

Deaton, A., and N. Cartwright (2016): "Understanding and Misunderstanding Randomized Controlled Trials," *NBER Working Paper*, (22595).

Deuchert, E., M. Huber, and M. Schelker (2016): "Mediation Analysis based on Differences in Differences," *working paper*.

Ferracci, M., G. Jolivet, and G. J. van den Berg (2010): "Treatment Evaluation in the Case of Interactions within Markets," *IZA DP No. 4700*.

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

Forastiere, L., F. Mealli, and T. J. VanderWeele (2016): "Identification and Estimation of Causal Mechanisms in Clustered Encouragement Designs: Disentangling Bed Nets Using Bayesian Principal Stratification," *Journal of the American Statistical Association*, 111, 510–525.

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

Frölich, M., and M. Huber (2014): "Direct and Indirect Treatment Effects: Causal Chains and Mediation Analysis with Instrumental Variables," *IZA Discussion Paper*, 8280.

Frölich, M., and K. Michaelowa (2011): "Peer effects and textbooks in African primary education," *Labour Economics*, 18(4), 474 – 486.

Graham, B. (2008): "Identifying Social Interactions Through Conditional Variance Restrictions," *Econometrica*, 76, 643–660.

Graham, B., G. W. Imbens, and G. Ridder (2010): "Measuring the Effects of Segregation in the Presence of Social Spillovers: A Nonparametric Approach," NBER Working Paper No. 16499.

Hahn, J., P. Todd, and W. van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

Heckman, J., L. Lochner, and C. Taber (1998): "General Equilibrium Treatment Effects: A Study of Tution Policy," *NBER Working Paper No. 6426*.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Hong, G. (2010): "Ratio of mediator probability weighting for estimating natural direct and indirect effects," in *Proceedings of the American Statistical Association, Biometrics Section*, pp. 2401–2415. Alexandria, VA: American Statistical Association.

Hong, G., and S. W. Raudenbush (2006): "Evaluating Kindergarten Retention Policy," *Journal of the American Statistical Association*, 101, 901–910.

Huber, M. (2014): "Identifying causal mechanisms (primarily) based on inverse probability weighting," *Journal of Applied Econometrics*, 29, 920–943.

Hudgens, M. G., and M. E. Halloran (2008): "Toward Causal Inference With Interference," *Journal of the American Statistical Association*, 103, 832–842.

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

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

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

Lalive, R., and M. A. Cattaneo (2009): "Social Interactions and Schooling Decisions," *Review of Economics and Statistics*, 91, 457–477.

Lalive, R., C. Landais, and J. Zweimüller (2015): "Market Externalities of Large Unemployment Insurance Extension Programs," *American Economic Review*, 105(12), 3564–3596.

Lechner, M. (2009): "Sequential Causal Models for the Evaluation of Labor Market Programs," *Journal of Business and Economic Statistics*, 27, 71–83.

Lechner, M., and R. Miquel (2010): "Identification of the effects of dynamic treatments by sequential conditional independence assumptions," *Empirical Economics*, 39, 111–137.

Manski, C. F. (1993): "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60, 531–542.

Miguel, E., and M. Kremer (2004): "Worms: Identifying impacts on education and health in the presence of treatment externalities," *Econometrica*, 72, 159–217.

Moffitt, R. A. (2001): *In Social Dynamics*chap. Policy Interventions, Low-Level Equilibria, and Social Interactions., pp. 45–82. MIT Press, Cambridge.

Nimczik, J. (2016): "Job Mobility Networks and Endogenous Labor Markets," *mimeo*.

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.

Robins, J. M. (1986): "A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect," *Mathematical Modelling*, 7, 1393–1512.

——— (1989): "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, ed. by L. Sechrest, H. Freeman, and A. Mulley, pp. 113–159. U.S. Public Health Service, Washington, DC.

——— (2003): "Semantics of causal DAG models and the identification of direct and indirect effects," in *In Highly Structured Stochastic Systems*, ed. by P. Green, N. Hjort, and S. Richardson, pp. 70–81, Oxford. Oxford University Press.

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

Robins, J. M., M. A. Hernan, and B. Brumback (2000): "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550–560.

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

——— (1990): "Formal mode of statistical inference for causal effects," *Journal of Statistical Planning and Inference*, 25, 279–292.

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

Sobel, M. E. (2006): "What Do Randomized Studies of Housing Mobility Demonstrate?," *Journal of the American Statistical Association*, 101, 1398–1407.

VanderWeele, T. J. (2008): "Simple relations between principal stratification and direct and indirect effects," *Statistics & Probability Letters*, 78, 2957–2962.

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

——— (2012): "Comments: Should Principal Stratification Be Used to Study Mediational Processes?," *Journal of Research on Educational Effectiveness*, 5(3), 245–249.

VANSTEELANDT, S., AND T. VANDERWEELE (2012): "Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions," *Biometrics*, 68, 1019–1027.

A Appendix: Proofs of Theorems

A.1 Proof of equation (17)

Under Assumptions 2, 4, and 7,

$$\begin{aligned}
\theta_{Z=1}(D(z)) &= \theta(D(z)) = E_{D(z)}[E[Y(1,d) - Y(0,d)|D(z) = d]] \\
&= E_{D|Z=z}[E[Y|Z=1,D] - E[Y|Z=0,D]] \\
&= E_{D|Z=z}\left[\frac{E[Y \cdot Z|D]}{\Pr(Z=1|D)} - \frac{E[Y \cdot (1-Z)|D]}{1-\Pr(Z=1|D)}\right] \\
&= E_{D|Z=z}\left[\frac{E[Y \cdot Z|D]}{\Pr(Z=1|D)} - \frac{E[Y \cdot (1-Z)|D]}{1-\Pr(Z=1|D)}\right] \\
&= E\left[\left(\frac{Y \cdot Z}{\Pr(Z=1|D)} - \frac{Y \cdot (1-Z)}{1-\Pr(Z=1|D)}\right) \cdot \frac{\Pr(Z=z|D)}{\Pr(Z=z)}\right].
\end{aligned}$$

The second equality follows from Assumptions 2 and 4, the fourth from Bayes' theorem and the fifth from the law of iterated expectations. Also note that $E_{A|B}[C]$ denotes the expectation of C over A conditional on B.

A.2 Proof of equations (20), (21), and (22)

Under Assumptions 8, 9, and 10(b),

$$\begin{aligned}
\delta_{Z=z,D=1}(z) &= E[Y(z,1) - Y(z,0)|Z=z, D=1] = E_{X|Z=z,D=1}[E[Y(z,1) - Y(z,0)|Z=z, D=1, X]] \\
&= E_{X|Z=z,D=1}[E[Y|Z=z,D=1,X] - E[Y|Z=z,D=0,X]], \\
&= E_{X|Z=z,D=1}\left[\frac{E[Y \cdot D|Z=z,X]}{\Pr(D=1|Z=z,X)} - \frac{E[Y \cdot (1-D)|Z=z,X]}{1-\Pr(D=1|Z=z,X)}\right] \\
&= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|Z=z,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|Z=z,X)}\right) \cdot \frac{\Pr(Z=z|X) \cdot \Pr(D=1|Z=z,X)}{\Pr(Z=z) \cdot \Pr(D=1|Z=z)}\right],
\end{aligned}$$

where the second equality follows from the law of iterated expectations, the third from Assumptions 8 and 9, the fourth from probability theory, and the fifth from the law of iterated expectations and Bayes' theorem.

Under Assumptions 8, 9, and 10(c),

$$\begin{aligned}
\theta_{Z=z,D=d}(d) &= E[Y(1,d)-Y(0,d)|Z=z,D=1] = E_{X|Z=z,D=d}[E[Y(1,d)-Y(0,d)|Z=z,D=1,X]] \\
&= E_{X|Z=z,D=d}[E[Y|Z=1,D=d,X]-E[Y|Z=0,D=d,X]] \\
&= E_{X|Z=z,D=d}\left[\frac{E[Y\cdot Z|D=d,X]}{\Pr(Z=1|D=d,X)}-\frac{E[Y\cdot(1-Z)|D=d,X]}{1-\Pr(Z=1|D=d,X)}\right] \\
&= E\left[\left(\frac{Y\cdot Z}{\Pr(Z=1|D=d,X)}-\frac{Y\cdot(1-Z)}{1-\Pr(Z=1|D=d,X)}\right)\cdot\frac{\Pr(Z=z|X)\cdot\Pr(D=d|Z=z,X)}{\Pr(D=d)\cdot\Pr(Z=z|D=d)}\right],
\end{aligned}$$

where the second equality follows from the law of iterated expectations, the third from Assumptions 8 and 9, the fourth from probability theory, and the fifth from the law of iterated expectations and Bayes' theorem.

Under Assumptions 8, 9, and 10(d),

$$\begin{aligned}
\theta_{Z=1}(D(z)) &= E_{X|Z=1}[E_{D(z)|Z=1,X}[E[Y(1,d)-Y(0,d)|D(z)=d,X]]] \\
&= E_{X|Z=1}\left[E_{D|Z=z,X}\left[E[Y|Z=1,D,X]-E[Y|Z=0,D,X]\right]\right] \\
&= E_X\left[E_{D|X}\left[\left(\frac{E[Y\cdot Z|D,X]}{\Pr(Z=1|D,X)}-\frac{E[Y\cdot(1-Z)|D,X]}{1-\Pr(Z=1|D,X)}\right)\cdot\frac{\Pr(Z=z|D,X)}{\Pr(Z=z|X)}\right]\cdot\frac{\Pr(Z=1|X)}{\Pr(Z=1)}\right] \\
&= E\left[\left(\frac{Y\cdot Z}{\Pr(Z=1|D,X)}-\frac{Y\cdot(1-Z)}{1-\Pr(Z=1|D,X)}\right)\cdot\frac{\Pr(Z=z|D,X)}{\Pr(Z=z|X)}\cdot\frac{\Pr(Z=1|X)}{\Pr(Z=1)}\right],
\end{aligned}$$

where the first equality follows from the law of iterated expectations, the second from Assumptions 8 and 9, the third from probability theory, and the fourth from the law of iterated expectations and Bayes' theorem.

A.3 Proof of equations (25) and (27)

Under Assumptions 3, 12, and 13,

$$\begin{aligned}
& E[Y_1|Z=1,\mathcal{T}=n]-E[Y_0|Z=1,\mathcal{T}=n]-[E[Y_1|Z=0,\mathcal{T}=n]-E[Y_0|Z=0,\mathcal{T}=n]] \\
= \;& E[Y_1(1,0)|Z=1,\mathcal{T}=n]-E[Y_0(1,0)|Z=1,\mathcal{T}=n] \\
- \;& [E[Y_1(0,0)|Z=0,\mathcal{T}=n]-E[Y_0(0,0)|Z=0,\mathcal{T}=n]] \\
= \;& E[Y_1(1,0)|Z=1,\mathcal{T}=n]-E[Y_0(0,0)|Z=1,\mathcal{T}=n] \\
- \;& [E[Y_1(0,0)|Z=0,\mathcal{T}=n]-E[Y_0(0,0)|Z=0,\mathcal{T}=n]] \\
= \;& E[Y_1(1,0)|Z=1,\mathcal{T}=n]-E[Y_0(0,0)|Z=1,\mathcal{T}=n] \\
- \;& [E[Y_1(0,0)|Z=1,\mathcal{T}=n]-E[Y_0(0,0)|Z=1,\mathcal{T}=n]] \\
= \;& E[Y_1(1,0)|Z=1,\mathcal{T}=n]-E[Y_1(0,0)|Z=1,\mathcal{T}=n]=\theta_{n,Z=1}(0),
\end{aligned}$$

where the first equality follows from the fact that never takers are identified by Assumption 3 and the observational rule ($Y|Z=z,\mathcal{T}=n$ corresponds to $Y(z,0)|\mathcal{T}=n$), the second from Assumption 13 ($Y_0(1,0)=Y_0(0,0)$), and the third from Assumption 12. The proof for the total effect on the compliers in treated regions ($\Delta_{c,Z=1}$) is analogous and therefore omitted.

Under Assumptions 3, 12, 13, and 14,

$$
\begin{aligned}
& E[Y_1|Z=2,\mathcal{T}=c] - E[Y_0|Z=2,\mathcal{T}=c] - [E[Y_1|Z=1,\mathcal{T}=c] - E[Y_0|Z=1,\mathcal{T}=c]] \\
= \;& E[Y_1(2,1)|Z=2,\mathcal{T}=c] - E[Y_0(2,1)|Z=2,\mathcal{T}=c] \\
- \;& [E[Y_1(1,1)|Z=1,\mathcal{T}=c] - E[Y_0(1,1)|Z=1,\mathcal{T}=c]] \\
= \;& E[Y_1(2,1)|Z=2,\mathcal{T}=c] - E[Y_0(0,0)|Z=2,\mathcal{T}=c] \\
- \;& [E[Y_1(1,1)|Z=1,\mathcal{T}=c] - E[Y_0(0,0)|Z=1,\mathcal{T}=c]] \\
= \;& E[Y_1(2,1)|Z=2,\mathcal{T}=c] - E[Y_0(0,0)|Z=2,\mathcal{T}=c] \\
- \;& [E[Y_1(1,1)|Z=1,\mathcal{T}=c] - E[Y_1(0,0)|Z=1,\mathcal{T}=c] \\
+ \;& E[Y_1(0,0)|Z=1,\mathcal{T}=c] - E[Y_0(0,0)|Z=1,\mathcal{T}=c]] \\
= \;& E[Y_1(2,1)|Z=2,\mathcal{T}=c] - E[Y_0(0,0)|Z=2,\mathcal{T}=c] \\
- \;& [E[Y_1(1,1)|Z=2,\mathcal{T}=c] - E[Y_1(0,0)|Z=2,\mathcal{T}=c] \\
+ \;& E[Y_1(0,0)|Z=2,\mathcal{T}=c] - E[Y_0(0,0)|Z=2,\mathcal{T}=c]] \\
= \;& E[Y_1(2,1)|Z=2,\mathcal{T}=c] - E[Y_1(1,1)|Z=1,\mathcal{T}=c] = \theta_{c,Z=2}(z'=2,z=1,d=1),
\end{aligned}
$$

where the first equality follows from the fact that compliers are identified by Assumption 3 and the observational rule ($Y|Z=z,\mathcal{T}=c$ corresponds to $Y(z,1)|\mathcal{T}=c$ for $z>0$), the second from Assumption 13 ($Y_0(1,0)=Y_0(0,0)$), the third from subtracting and adding $E[Y_1(0,0)|Z=1,\mathcal{T}=c]$, and the fourth from Assumptions 12 and 14.