Estimating Treatment Effects in Mover Designs*

Peter Hull†

October 18, 2017

Abstract

Researchers increasingly leverage quasi-experimental movement across regional or institutional treatments to estimate multiple causal effects. Often such "mover regressions" are motivated by additively-separable and constant-effect causal model; it is not clear, however, how to interpret regression estimates when these models are misspecified. I show that under standard parallel trends and stationarity assumptions, mover regressions with two treatments and two time periods identify a convex average of heterogeneous treatment effects, but that stronger restrictions on outcome variability are needed for causal interpretation when treatment is multi-valued. I propose instead a class of non-parametric estimators identifying a set of *mover average treatment effects* (MATEs) under weaker conditional parallel trend and effect homogeneity assumptions. I characterize the efficient estimator in this class, and derive specification tests from the model's overidentifying restrictions.

*I thank Alberto Abadie, Isaiah Andrews, Josh Angrist, Paul Goldsmith-Pinkham, and Chris Walters for valuable feedback. This draft is a work in progress and should not be cited without prior approval. Comments are welcome; all errors are my own.

†The University of Chicago and Microsoft Research. Email: hull@uchicago.edu; web: http://peterhull.net.

1 Introduction

The growing availability of rich longitudinal data has broadened the scope for causal inference in economics. Researchers increasingly exploit variation in an individual's institutional choices over time – such as the firm they work for, the city they live in, the teacher they learn from, or the doctor they are treated by – in order to estimate the average effects of a large number of these treatments simultaneously.[1] Typically these estimates are obtained from a linear two-way fixed effects regression that is motivated by a stationary, additive, and constant-effect causal model (e.g. Abowd et al. (1999)). It is not clear, however, what such "mover regressions" capture under likely misspecifications of this model, including outcome persistence, time-varying shocks, and heterogeneous treatment effects.

In this paper, I consider the causal content of quasi-experimental mover designs under weaker statistical assumptions. I first show that in the simplest case of a binary treatment, two time periods, and no additional controls, a stationarity and parallel trends restriction are sufficient for two-way fixed effect regressions to identify a convex average of causal effects. This finding links mover analyses to more familiar difference-in-differences designs, but turns out to not generalize to mover quasi-experiments with multiple institutional treatments. Indeed, the most common many-treatment mover regressions are not causally interpretable under stationarity and parallel trends alone, but require additional strong restrictions on outcome variability across both treatments and time.

Alternative estimators, however, retain the flexibility of the difference-in-differences approach in more complex settings. I first develop a two-step non-parametric weighting procedure for estimating a set of causal parameters – termed *mover average treatment effects* (MATEs) – in settings with a binary treatment and a conditional-on-covariates parallel trends assumption. This estimator builds on Abadie (2005), who proposes a similar procedure for difference-in-differences designs, and unlike with classical mover regressions does not require outcome stationarity. I then extend this result to a general class of estimators in multi-treatment, multi-period settings that leverage conditional-on-covariate restrictions on both outcome trends and mover treatment effects. The key effect homogeneity assumption in multi-treatment experiments is weaker than the usual constant effects restriction, yet is still testable via a large set of overidentifying restrictions. As in the binary treatment case, no restriction on outcome persistence is required. I characterize the efficient MATE estimator within this class, given a set of first-step propensity scores and argue that, relative to the complexity of high-dimensional fixed effects regressions, this estimator is likely to be easily computed.

This analysis contributes to a small but growing literature relaxing the strong assumptions of canonical movers models, typically as applied to matched worker-firm panels. Abowd et al. (2015), for example, propose tests of the additive constant-effects model and develop a Bayesian approach to capture non-random worker movement, while Hagedorn et al. (2017) leverage structural assumptions and long-run time series

[1] Recent examples include Bronnenberg et al. (2012), Card et al. (2013), Jackson (2013), Chetty et al. (2014), Bloom et al. (2015), Finkelstein et al. (2016), Sacarny (2016), Chetty and Hendren (2017), Finkelstein et al. (2017), and Allcott et al. (2017).

variation to estimate worker and firm effect ranks. Most recently, Bonhomme et al. (2017) show how to accomodate discrete heterogeneity and Markovian movement patterns with certain forms of endogeneity. To my knowledge no paper has yet to study mover designs within a treatment effects framework.[2]

As suggested above, this paper can also be thought to generalize classical and recent approaches to difference-in-differences estimation – including Ashenfelter and Card (1985), Heckman et al. (1997, 1998), Abadie (2005), Athey and Imbens (2006) and Imai and Kim (2016) – to settings with multiple unordered treatments. More generally this work builds on the canonical panel data literature (Chamberlain, 1980, 1982, 1984; Manski, 1987; Honore, 1992; Arellano, 2003) by allowing for certain forms of nonlinearity, non-additivity, and heterogeneity in causal response. Notably, the key parallel trends assumption considered here is weaker than the "time ignorability" restriction typically used by this literature, most recently in Hahn (2001), Wooldridge (2005), and Chernozhukov et al. (2013).[3] The dynamic potential outcomes framework also allows for the kinds of non-stationarity and outcome persistence often ruled out implicitly or explicitly in this panel literature (Robins et al., 2000; Imai and Kim, 2016). As with many treatment effects approaches, a cost of this increased generality is a narrower set of identified causal estimands: namely, average treatment effects for the subset of individuals who move into the treatment at a particular point in time, or loosely per Imbens and Angrist (1994), the mover experiment's "compliers."

For simplicity, the next sections outlines the main approach in two-period mover designs. Section 3 then extends the theory to the general multi-period case, and section 4 concludes. Future drafts of this paper will apply the theory to several recent examples of mover designs, in order to demonstrate computational issues and empirical relevance.

2 Two-period Mover Designs

For a panel of individuals i in T time periods $t = 0, \ldots, T-1$, we observe an outcome Y_{it}, a vector of covariates X_{it}, and an individual's choice $J_{it} \in \{0, \ldots, J-1\}$ among J possible treatments. Let $D_{ijt} = \mathbf{1}[J_{it} = j]$ be an indicator for individual i choosing treatment j in time t. Conventional mover analyses estimate the ordinary least squares regression

$$Y_{it} = \alpha_i + \tau_t + \sum_{j \neq 0} \beta_j D_{ijt} + X_{it}'\gamma + \epsilon_{it}. \tag{1}$$

Here α_i and τ_t denote the two-way individual and time fixed effects, respectively, while the coefficient β_j is meant to capture the effect of treatment j relative to the omitted treatment category 0. I refer to equation (1) as the *mover regression.*[4]

[2] De Chaisemartin and D'Haultfoeuille (2016) discuss identification of local average treatment effects in instrumental variable models with two-way fixed effects; although in a treatment effects context, their motivation, setting, and analysis are distinct.

[3] Graham and Powell (2012) consider an alternative approach to panel identification with continuous treatment variables; typical mover designs study choice among discrete unordered alternatives, so this theory is less relevant here.

[4] Sometimes the mover regression is written $Y_{it} = \alpha_i + \tau_t + \beta_{J_{it}} + X_{it}'\gamma + \epsilon_{it}$: see, e.g., Card et al. (2013). One treatment category is always omitted from estimation in practice as typically only relative treatment coefficients are identified.

To characterize the causal interpretation of mover regressions, I use the dynamic potential outcomes framework of Robins (1986, 1997), which in turn builds on Rubin (1974). Let $Y_{it}^{k\to j}$ denote the potential outcome of individual i in time t if she were to select treatment j after choosing a treatment path of $k = (k_0, \ldots, k_{t-1})'$. When not ambiguous I write $Y_{it}^{j} \equiv Y_{it}^{(J_{i0},\ldots,J_{i,t-1})'\to j}$ as the time-t potential outcome of individual i, given the set of her actual previous treatment choices J_{is}. Point-in-time treatment effects relative to the omitted category are written $Y_{it}^{j} - Y_{it}^{0}$, and realized outcomes can be expressed, with $\overline{0} = (0, \ldots, 0)'$,

$$\begin{aligned} Y_{it} =& Y_{it}^{0} + \sum_{j\neq 0}(Y_{it}^{j} - Y_{it}^{0})D_{ijt} \\ =& Y_{it}^{\overline{0}\to 0} + \sum_{j\neq 0}(Y_{it}^{\overline{0}\to j} - Y_{it}^{\overline{0}\to 0})D_{ijt} \\ &+ \sum_{k\neq 0}\left(Y_{it}^{k\to 0} - Y_{it}^{\overline{0}\to 0} + \sum_{j\neq 0}(Y_{it}^{k\to j} - Y_{it}^{\overline{0}\to j} - (Y_{it}^{k\to 0} - Y_{it}^{\overline{0}\to 0}))D_{ijt}\right)\prod_{s<t} D_{ik_s s}. \end{aligned} \tag{2}$$

A comparison of equations (1) and (2) suggests one sufficient set of conditions for mover regression estimates to be causally interpretable:

Assumption SO (*Static outcomes*): For all j, t, and k, $P(Y_{it}^{k\to j} = Y_{it}^{j}) = 1$

Assumption CE (*Constant effects*): For all j and t, there exists $\overline{\beta}^{j}$ such that $P(Y_{it}^{\overline{0}\to j} - Y_{it}^{\overline{0}\to 0} = \overline{\beta}^{j}) = 1$

Assumption CO (*Conditional orthogonality*): For all j and t, $E[D_{ijt}\widetilde{\epsilon_{it}}] = 0$, where $\widetilde{\epsilon_{it}}$ denotes the population residual from projecting $Y_{it}^{\overline{0}\to 0}$ on X_{it} and individual and time effects

Potential outcomes are stationary under Assumption SO, in that they only depend on contemporaneous treatment; when this is the case the last term of equation (2) is zero. If furthermore the period-specific treatment effects are constant across individuals (Assumption CE), we may write $Y_{it} = Y_{it}^{0} + \sum_{j\neq 0}\overline{\beta}^{j}D_{ijt}$, so that if potential outcome heterogeneity is also conditionally orthogonal to treatment choice (Assumption CO), the β_j in equation (1) identify the constant causal effects $\overline{\beta}^{j}$. These conditions are straightforward to state mathematically and are often motivated in practice by an underlying economic model (e.g. Finkelstein et al. (2016)). Nevertheless, they are also strong: in practice, a researcher may be reluctant to completely rule out treatment effect persistence and heterogeneity, or may have difficulty assessing the appropriateness of Assumption CO in different settings. Previous theoretical work relaxes some, but not all of these assumptions, while retaining their mathematical flavor.[5] The remainder of this section instead considers what regression (1) identifies more generally; I then propose new estimators that are consistent for a set of causal effects when Assumptions SO, CE, and CO are replaced by weaker, and likely familiar restrictions. To this end I first consider the simplest possible setting, without covariates and with only two treatment states and two time periods, before relaxing these simplifications.

[5]For example Wooldridge (2005) and Chernozhukov et al. (2013) replace Assumptions CE and CO by a weaker "time ignorability" condition. As discussed below, this implies the key parallel trends assumption I consider, but is not equivalent.

2.1 Binary Treatment in Two Periods

Note that when $T = 2$, the mover regression (1) can be equivalently estimated by first-differences:

$$\Delta Y_i = \tau + \sum_{j \neq 0} \beta_j \Delta D_{ij} + \Delta X_i' \gamma + \Delta \epsilon_i, \tag{3}$$

where $\Delta V_i = V_{i1} - V_{i0}$ denotes the first-difference operator applied to variable V_{it} and $\tau = \tau_1 - \tau_0$. When furthermore treatment is binary ($J = 2$) and there are no added covariates ($X_{it} = 0$), we have the following algebraic expression for the mover regression coefficient:

Lemma 1: If $T = J = 2$ and $X_{it} = 0$, then the mover regression coefficient equals

$$\begin{aligned} \beta_1 =& (E[\Delta Y_i \mid \Delta D_{i1} = 1] - E[\Delta Y_i \mid \Delta D_{i1} = 0])\omega \\ &+ (E[\Delta Y_i \mid \Delta D_{i1} = 0] - E[\Delta Y_i \mid \Delta D_{i1} = -1])(1 - \omega), \end{aligned} \tag{4}$$

where $\omega \in [0, 1]$.

The derivation of Lemma 1, contained in the appendix along with all other proofs, uses omitted-variables bias algebra to write β_1 as a linear combination of the coefficients of a saturated model for $E[\Delta Y_i \mid \Delta D_{i1}]$, which identifies mean outcome growth among those with $\Delta D_{i1} = -1$, $\Delta D_{i1} = 0$, and $\Delta D_{i1} = 1$. Throughout this section I refer to individuals with $\Delta D_{ij} \neq 0$ for some j as *movers* and those with $\Delta D_{ij} = 0$ for all j as *stayers*.

Using Lemma 1, it is straightforward to verify that a stationarity assumption, along with a parallel trends assumption similar to those often used in difference-in-differences analyses, renders the simplest mover coefficient causally interpretable:

Proposition 1: If $T = J = 2$, $X_{it} = 0$, and potential outcomes satisfy

$$E[Y_{i1}^{1\to 0} \mid \Delta D_{i1} = -1] = E[Y_{i1}^{0\to 0} \mid \Delta D_{i1} = -1], \tag{5}$$

$$E[Y_{i1}^{0\to 1} \mid \Delta D_{i1} = 1] = E[Y_{i1}^{1\to 1} \mid \Delta D_{i1} = 1], \tag{6}$$

and

$$\begin{aligned} E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid D_{ij0} = D_{ij1} = 1] =& E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid D_{ij0} = D_{ik1} = 1] \\ =& E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid D_{ik0} = D_{ij1} = 1] \end{aligned} \tag{7}$$

for all $j \in \{0, 1\}$ and $k \neq j$, then the mover regression coefficient identifies

$$\beta_1 = \sum_{t \in \{0,1\}} \sum_{d \in \{-1,1\}} E[Y_{it}^1 - Y_{it}^0 \mid \Delta D_{i1} = d]\, \omega_{td}, \tag{8}$$

where $\omega_{td} \geq 0$ and $\sum_{t \in \{0,1\}} \sum_{d \in \{-1,1\}} \omega_{td} = 1$. Moreover, this weighting scheme is identified.

Proposition 1 shows that the mover regression coefficient β_1 identifies a convex combination of average

treatment effects for movers across the two time periods under two conditions. First, individuals moving into each treatment at time 1 must on average have the same outcome as if they had always been there (a stationarity assumption). Second, conditional on an individual being in treatment j at any point, the average outcomes for different groups of movers and stayers follow the same paths in the absence of a move (a parallel trends assumption). The weights ω_{td} are shown in the appendix to depend on the distribution of types of movers and stayers. Intuitively, if there are no movers with $\Delta D_{i1} = -1$ then β_1 is weighted average of causal effects across the two periods among those with $\Delta D_{i1} = 1$, and vice-versa. How much weight is placed on effects from time 0 versus time 1 moreover depends on the proportion of stayers in the treated and untreated states: at the extremes if there are no individuals with $D_{id0} = D_{id1} = 0$, for either $d = 0$ or $d = 1$, then the mover regression weights together time-d effects for movers with $\Delta D_{i1} = 1$ and time-$(1-d)$ effects for movers with $\Delta D_{i1} = -1$. Thus, if no individual is initially treated, β_1 estimates $E[Y_{i1}^1 - Y_{i1}^0 \mid D_{i1} = 1]$, the average effect on individuals who become treated in time 1, as in a standard difference-in-differences design.

Although causally interpretable, this result shows that mover regression coefficients do not, in general, identify a simple time-t mover average treatment effect, defined here as

$$MATE_{jt} = E[Y_{it}^j - Y_{it}^0 \mid \Delta D_{ij} \neq 0]. \tag{9}$$

Even in the simplest setting with only two treatments, two time periods, and no covariates, two mover experiments with the same joint distribution of causal effects and treatment choices (and thus the same MATEs) may produce different regression coefficients depending on the marginal distribution of treatment. Researchers interested in the external validity of their mover regression estimates, or in comparing estimates across different experiments, may view this as an important limitation of the standard regression approach.[6]

Nevertheless, it is straightforward to verify that the two MATEs for treatment 1 are, in fact, identified under the assumption used in Proposition 1. As shown in the appendix, the difference in outcome growth rates between stayers with $D_{i00} = D_{i01} = 1$ and movers with $\Delta D_{i1} = -1$ identifies the time-0 treatment effect of the latter group when the stationarity and parallel trends assumptions hold:

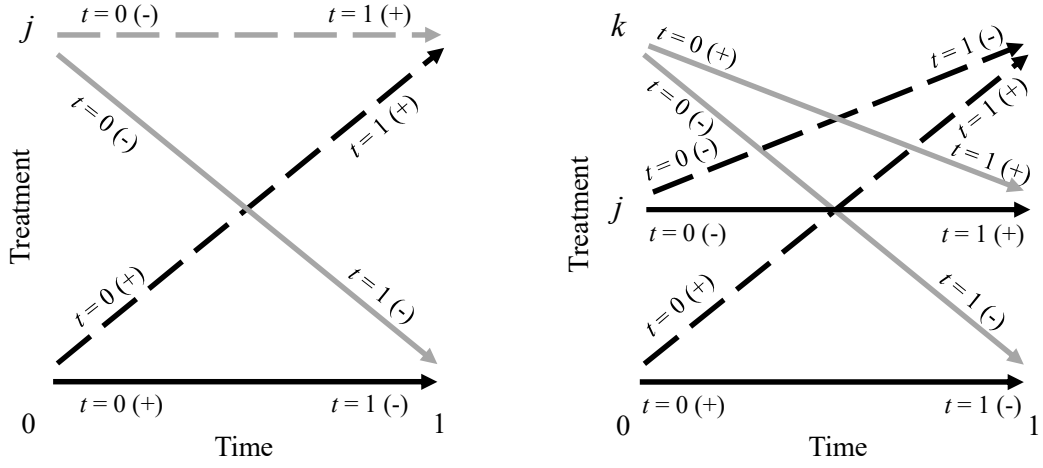
$$E[\Delta Y_i \mid \Delta D_{i1} = 0, D_{i00} = 1] - E[\Delta Y_i \mid \Delta D_{i1} = -1] = E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} = -1] \tag{10}$$

Moreover, subtracting the outcome growth of the other movers with $\Delta D_{i1} = 1$ from stayers with $D_{i10} = D_{i11} = 1$ identifies the time-0 treatment effect for the former group:

$$E[\Delta Y_i \mid \Delta D_{i1} = 1] - E[\Delta Y_i \mid \Delta D_{i1} = 0, D_{i10} = 1] = E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} = 1] \tag{11}$$

A researcher could thus compute these growth differentials and average them together by the proportion of movers of each type to non-parametrically estimate $MATE_{10}$. Similarly, combining the excess outcome growth of $\Delta D_{i1} = -1$ movers relative to the $D_{i10} = D_{i11} = 1$ stayers with the difference in outcome growth

[6] Yitzhaki (1996) argues this point for regression-weighted averages of causal parameters in general; see also Angrist (1998).



(a) Requiring Assumption CPT

(b) Requiring Assumptions CPT and CEH

Figure 1: Mover and stayer group comparisons identifying time-t treatment effects

Notes: This figure illustrates the groups of movers and stayers identifying the time-t effects of treatment j. In each panel, time-0 effects are identified by outcome growth contrasts within the dashed-line group and within the solid-line group, while time-1 effects are identified by outcome growth contrasts within the light-colored group and the dark-colored group. The time-specific notes on each line indicate whether the outcome growth for that subgroup is added or subtracted. Panel (a) shows comparisons involving movers between treatment j and 0, for which only Assumption CPT is required, while panel (b) shows comparisons involving other movers which require both Assumptions CPT and CEH.

between $\Delta D_{i1} = 1$ movers and $D_{i00} = D_{i01} = 1$ stayers will consistently estimate the time-1 mover average treatment effect, $MATE_{11}$.

These comparisons between various mover and stayer types are summarized in Figure 1(a). Outcome growth contrasts within the dashed-line group and within the solid-line group identify the time-0 MATEs, while comparisons within the dark-colored group and within the light-colored group identify the time-1 MATEs. Interestingly, as is apparent from the proof to Proposition 1, only the parallel trends assumption is used for identifying time-1 MATEs, whereas stationarity is also needed for estimating time-0 MATEs. Intuition for this also comes from Figure 1(a): time-0 outcomes are comparable and can thus be "differenced off" within the light-colored and dark-colored comparisons under Assumption CPT, but time-1 outcomes within the solid and dashed line groups are not comparable if potential outcomes are not stationary.

One may formalize the identification argument illustrated in Figure 1(a) and extend it to use conditional versions of the assumptions used in Proposition 1, which may be weaker than unconditional stationarity and parallel trends in practice. Namely, with X_i denoting a vector of time-invariant controls (including, for example, some subset of the X_{i0} and X_{i1} vectors from the mover regression), the key parallel trends assumption is

Assumption CPT (*Conditional parallel trends*): For treatment j, $E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid D_{ij0} = D_{ij1} = 1, X_i] = E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid D_{ik0} = D_{ik1} = 1, X_i] = E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid D_{ik0} = D_{ij1} = 1, X_i]$, for all $k \neq j$, provided these are well-defined.

Under Assumption CPT, the average treatment-j outcomes for different types of movers at j would have followed parallel paths if not for the move, given the controls. In many settings it may be plausible that an individual's treatment selection is only driven by potential outcome dynamics through a set of contemporaneous or lagged observables, as with the famous "Ashenfelter dip" of income for those entering job training programs (Ashenfelter, 1978; Ashenfelter and Card, 1985; Abadie, 2005). In practice many mover papers conduct informal tests of this assumption by studying regression-adjusted pre-trends (e.g. Finkelstein et al. (2016)). Clearly, the second assumption in Proposition 1 is a special case of Assumption CPT, for which $X_i = 0$; it is also straightforward to verify that the usual "time ignorability" assumptions underlying classical panel identification imply Assumption CPT, but are not implied by it.[7]

Given this, we have the following identification result for binary treatment mover designs:

Proposition 2: If $T = J = 2$ and, for $t = 1$,

1. Assumption CPT holds for treatments 0 and 1
2. $P(D_{i00} = D_{i01} = 1 \mid X_i) > 0$ for all values of X_i such that $P(D_{i1t} = D_{i0,1-t} \mid X_i) > 0$
3. $P(D_{i10} = D_{i11} = 1 \mid X_i) > 0$ for all values of X_i such that $P(D_{i0t} = D_{i1,1-t} \mid X_i) > 0$,

then

$$MATE_{1t} = E[\Delta Y_i(\kappa_{i1t} + \lambda_{i1t})], \tag{12}$$

where

$$\kappa_{ijt} = (-1)^t D_{i0,1-t} \frac{E[D_{ijt} D_{i0,1-t} \mid X_i] - D_{ijt} E[D_{i0,1-t} \mid X_i]}{E[D_{i00} D_{i01} \mid X_i] P(\Delta D_{ij} \neq 0)} \tag{13}$$

and

$$\lambda_{ijt} = (-1)^t D_{ij,1-t} \frac{E[D_{ijt} D_{ij,1-t} \mid X_i] - D_{ijt} E[D_{ij,1-t} \mid X_i]}{E[D_{ij0} D_{ij1} \mid X_i] P(\Delta D_{ij} \neq 0)}. \tag{14}$$

If the above conditions also hold for $t = 0$ and, moreover, we have

$$E[Y_{i1}^{1\to 0} \mid \Delta D_{i1} = -1, X_i] = E[Y_{i1}^{0\to 0} \mid \Delta D_{i1} = -1, X_i] \tag{15}$$

$$\text{and} \quad E[Y_{i1}^{0\to 1} \mid \Delta D_{i1} = 1, X_i] = E[Y_{i1}^{1\to 1} \mid \Delta D_{i1} = 1, X_i], \tag{16}$$

then equations (12)-(14) also hold for $t = 0$.

Proposition 2 shows that with two periods, binary treatment, and conditional parallel trends, the period-1 mover average treatment effect is identified by certain weighted averages of outcome growth ΔY_i. If a conditional stationarity restriction (equations (15) and (16)) also holds, the time-0 MATE is similarly identified. Here the weights κ_{ijt} and λ_{ijt} depend on the distribution of moves and non-moves given the

[7] For example, Chernozhukov et al. (2013) consider specifications of the form $Y_{it}^{k\to j} = g_0(j, \alpha_i, \epsilon_{it})$, where the distribution of ϵ_{it} does not depend on t given α_i and leads and lags of J_{it}. Thus $E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid J_{i0}, J_{i1}] = 0$, satisfying assumption CPT.

controls, as summarized by the mover "propensity scores" $E[D_{ij0}D_{ik1} \mid X_i]$, and are shown in the appendix proof to replicate and average together the conditional-on-X_i comparisons between mover and stayer groups illustrated in Figure 1(a). The second and third conditions in Proposition 2 ensure that the data are sufficiently rich for each of these comparisons to be feasible.

Proposition 2 can be thought to generalize previous approaches to difference-in-difference treatment effect estimation. Namely, if $P(\Delta D_{i1} = -1 \mid X_i) = 0$ for all values of X_i, then κ_{ij1} will be zero, and the weighting scheme identifying $MATE_{11}$ will coincide with that of Abadie (2005), comparing outcome growth across those with $\Delta D_{i1} = 1$ to those with $D_{i00} = D_{i01} = 1$. If, conversely, $P(\Delta D_{i1} = 1 \mid X_i) = 0$, the roles of κ_{ij1} and λ_{ij1} are reversed, and $MATE_{11}$ will be identified by an analogous comparison of the outcome growth of movers with $\Delta D_{i1} = -1$ and stayers with $\Delta D_{i10} = D_{i11} = 1$. In general, the weighting scheme in Proposition 2 combines average causal effects from both of these difference-in-difference quasi-experiments.

As in Abadie (2005), Proposition 2 suggests a straightforward two-step estimator for mover average treatment effects when treatment and time are both binary. In a first step, a researcher computes the sample proportion of movers, $\widehat{P}(\Delta D_{i1} \neq 0)$, and forms non-parametric series approximations to three conditional mean functions, $E[D_{i10} \mid X_i]$, $E[D_{i11} \mid X_i]$, and $E[D_{i10}D_{i11} \mid X_i]$. She then forms the sample analogue of equation (12), using the fact that $\widehat{E}[D_{i0t} \mid X_i] = 1 - \widehat{E}[D_{i1t} \mid X_i]$ for each t, and similar identities.[8] When these approximations are consistent, so too will be the second-step weighting estimator by Proposition 2. Inference on MATEs estimated in this way follows from the asymptotic theory of Andrews (1991) and others.

2.2 Multi-valued Treatment

In a typical mover design, researchers estimate regressions with many different treatment choices.[9] When $J > 2$, it is straightforward to see how Proposition 2 may be used to estimate average effects for particular subsets of movers for each treatment j: namely, those who move between j and the omitted treatment 0. This is because one can always mimic the binary treatment setting by restricting observations to those individuals who move from 0 to j (with $D_{i00} = D_{ij1} = 1$) or from j to 0 (with $D_{ij0} = D_{i01} = 1$), along with the corresponding stayers in both states (with $D_{ij0} = D_{ij1} = 1$ or $D_{i00} = D_{i01} = 1$). One can then construct the weights from Proposition 2 within this subsample, leading to the identification of a well-specified average causal effect. The following result formalizes this intuition:

Corollary to Proposition 2: For some j, let $M_{ij} = D_{i00}D_{ij1} + D_{ij0}D_{i01}$ and $S_{ij} = D_{ij0}D_{ij1} + D_{i00}D_{i01}$. If $T = 2$ and three conditions in Proposition 2 hold for some t and for treatments 0 and j (in place of 1), with $M_{ij} + S_{ij}$ included in X_i, then the average time-t effect of treatment j for movers between 0

[8] Namely, $\widehat{E}[D_{i10}D_{i01} \mid X_i] = \widehat{E}[D_{i10} \mid X_i] - \widehat{E}[D_{i10}D_{i11} \mid X_i]$, $\widehat{E}[D_{i00}D_{i11} \mid X_i] = \widehat{E}[D_{i11} \mid X_i] - \widehat{E}[D_{i10}D_{i11} \mid X_i]$, and $\widehat{E}[D_{i00}D_{i01} \mid X_i] = 1 - \widehat{E}[D_{i10} \mid X_i] - \widehat{E}[D_{i11} \mid X_i] - \widehat{E}[D_{i10}D_{i11} \mid X_i]$.

[9] For example, Finkelstein et al. (2016) estimate models with 305 medical market fixed effects, while Card et al. (2013) study effects from over a million German firms.

and j is given by

$$E[Y_{it}^{j} - Y_{it}^{0} \mid \Delta D_{ij} \neq 0, \Delta D_{i0} \neq 0] = E\left[\Delta Y_i \frac{(\kappa_{ijt} + \lambda_{ijt})(M_{ij} + S_{ij})}{P(M_{ij} = 1 \mid \Delta D_{ij} \neq 0)}\right] \tag{17}$$

where κ_{ijt} and λ_{ijt} are as in Proposition 2.

Equation (17) scales the weighting scheme from Proposition 2 by an additional term, ensuring that it only uses observations from the desired subpopulation of movers (with $M_{ij} = 1$) and stayers (with $S_{ij} = 1$) between treatment 0 and j. Note that when treatment is binary, $\Delta D_{ij} = -\Delta D_{i0}$ so that $E[Y_{it}^{j} - Y_{it}^{0} \mid \Delta D_{ij} \neq 0, \Delta D_{i0} \neq 0] = MATE_{jt}$; furthermore $M_{ij} = \mathbf{1}[\Delta D_{ij} \neq 0] = 1 - S_{ij}$, so the weighting scheme reduces to that of Proposition 2.

This result offers a straightforward way to recover well-defined causal effects from multi-valued treatment mover designs under Assumption CPT alone. Nevertheless, its usefulness is likely to be limited by the fact that it restricts attention to a specific, and perhaps unrepresentative subset of the mobile population. Furthermore, by making such a restriction it offers no way to compare effects across multiple treatments for the same set of movers, which is often of interest in mover designs. I next consider multi-valued treatment extensions to Proposition 2 that overcome these limitations.

We start, as before, by considering what simple mover regressions identify in the multi-treatment case. Unlike with binary treatment, however, this exercise yields a very different, and more negative, initial conclusion:

Proposition 3: When $J > 2$, mover regression coefficients do not in general estimate weighted averages of individual treatment effects under stationarity and parallel trend assumptions alone.

The proof of Proposition 3 shows that even in the simplest case of $T = 2$ and $X_{it} = 0$, the multi-valued treatment coefficients in equation (1) identify a linear combination of mover treatment effects *and* a set of non-causal terms, even under the parallel trends and stationarity assumptions, CPT and SO. The latter terms are of the form

$$E[Y_{i1}^{j} - Y_{i0}^{j} \mid D_{ij0} = D_{ij1} = 1] - E[Y_{i1}^{k} - Y_{i0}^{k} \mid D_{ik0} = D_{ik1} = 1] \tag{18}$$

and

$$E[Y_{it}^{j} - Y_{it}^{0} \mid D_{i00} = D_{ij1} = 1] + E[Y_{is}^{k} - Y_{is}^{j} \mid D_{ij0} = D_{ik1} = 1] - E[Y_{iu}^{k} - Y_{iu}^{0} \mid D_{i00} = D_{ik1} = 1], \tag{19}$$

for some $j, k \neq 0$ and $j \neq k$, and capture 1) the difference in outcome trends among stayers in different treatment states and 2) combinations of causal effects for different groups of movers at different treatments, respectively. In general, these non-causal terms are non-zero even under strong stationarity and parallel trends assumptions and given non-zero weight in the regression, rendering the mover coefficients causally uninterpretable. Each term is, however, equal to zero in the kind of additively-separable and constant effect models conventionally motivating mover designs, or directly by Assumptions SO, CE, and CO.

Also unlike the binary treatment case, it is straightforward to verify that the average treatment-j effects for individuals moving between treatment j and any $k \neq 0$ are not identified by parallel trends and stationarity alone. Intuitively, as these individuals are never observed in the omitted treatment category 0, no stayer comparison can directly reveals their treatment effects $Y_{it}^j - Y_{it}^0$. Nevertheless one can estimate the average causal effects from switching such individuals between treatment k and j (which I refer to as average "marginal" effects across treatments); this is given simply by the comparison of their outcome growth to that of stayers at treatment k. Thus if average treatment-k effects are comparable between these movers and other individuals moving between k and 0, we can indirectly estimate treatment-j effects for the desired mover subpopulation by combining two other causal parameters. As a stark example, note that under the constant-effect Assumption CE,

$$\underbrace{E[Y_{it}^k - Y_{it}^0 \mid D_{i00} = D_{ik1} = 1]}_{\text{Avg. treatment-}k\text{ effect for }0 \to k\text{ movers}} - \underbrace{E[Y_{it}^k - Y_{it}^j \mid D_{ij0} = D_{ik1} = 1]}_{\text{Avg. marginal effect for }j \to k\text{ movers}} = \overline{\beta}^k - (\overline{\beta}^k - \overline{\beta}^j)) = \overline{\beta}^j$$
$$= \underbrace{E[Y_{it}^j - Y_{it}^0 \mid D_{ij0} = D_{ik1} = 1]}_{\text{Avg. treatment-}j\text{ effect for }j \to k\text{ movers}}. \quad (20)$$

Rather of ruling out all treatment effect heterogeneity, however, MATE identification in the multi-valued treatment case can be achieved by a weaker assumption of mean-independence of treatment effects across certain subpopulations of movers, conditional on the time period and additional covariates:

Assumption CEH (*Conditional effect homogeneity*): For time period t and treatment j, the mean effects $E[Y_{it}^j - Y_{it}^0 \mid \Delta D_{ij} = d, \Delta D_{ik} = -d, X_i]$ do not depend on $k \neq j$, for all $d \neq 0$, provided they are well-defined.

Restrictions on the *conditional* heterogeneity of causal parameters of this kind are increasingly deployed in the treatment effects literature, particularly for issues of external validity within and across quasi-experimental designs (Angrist and Fernandez-Val, 2013; Angrist and Rokkanen, 2015; Hull, 2015). Here Assumption CEH states that systematic differences in treatment effects across certain movers with different origins or destinations are driven only by a set of contemporaneous or lagged observables contained in X_i. In practice researchers often gauge effect homogeneity in mover designs by informal tests of regression-adjusted trend symmetry across different mover types (e.g. Card et al. (2013)).

Given this restriction, we have the following result in the multi-valued treatment setting:

Proposition 4: If $T = 2$ and, for some j and $t = 1$,

1. Assumption CPT holds for 0 and j, and Assumption CEH holds for t and all $k \neq j$
2. $P(D_{i0t} = D_{ik,1-t} = 1 \mid X_i) > 0$ and $P(D_{ik0} = D_{ik1} = 1 \mid X_i) > 0$ for all $k \neq j$ and values of X_i for which $P(D_{ijt} = D_{ik,1-t} = 1 \mid X_i) > 0$
3. $P(D_{ikt} = D_{i0,1-t} = 1 \mid X_i) > 0$, $P(D_{ij0} = D_{ij1} = 1 \mid X_i) > 0$, and $P(D_{i00} = D_{i01} = 1 \mid X_i) > 0$ for all $k \neq j$ and values of X_i for which $P(D_{ikt} = D_{ij,1-t} = 1 \mid X_i) > 0$,

then

$$MATE_{jt} = E\left[\Delta Y_i \sum_{k \neq j} (\kappa_{ijt}^k + \lambda_{ijt}^k)\right], \tag{21}$$

where

$$\kappa_{ijt}^k = (-1)^t D_{ik,1-t} \frac{D_{i0t} E[D_{ijt} D_{ik,1-t} \mid X_i] - D_{ijt} E[D_{i0t} D_{ik,1-t} \mid X_i]}{E[D_{i0t} D_{ik,1-t} \mid X_i] P(\Delta D_{ij} \neq 0)} \tag{22}$$

and

$$\begin{aligned} \lambda_{ijt}^k =& (-1)^t D_{ij,1-t} \frac{D_{ikt} E[D_{ij0} D_{ij1} \mid X_i] - D_{ijt} E[D_{ikt} D_{ij,1-t} \mid X_i]}{E[D_{ij0} D_{ij1} \mid X_i] P(\Delta D_{ij} \neq 0)} \\ &+ (-1)^t D_{i0,1-t} \frac{D_{i0t} E[D_{ikt} D_{i0,1-t} \mid X_i] - D_{ikt} E[D_{i00} D_{i01} \mid X_i]}{E[D_{ikt} D_{i0,1-t} \mid X_i] E[D_{i00} D_{i01} \mid X_i]} \frac{E[D_{ikt} D_{ij,1-t} \mid X_i]}{P(\Delta D_{ij} \neq 0)}. \end{aligned} \tag{23}$$

If the above conditions also hold for $t = 0$ and, moreover, the stationarity conditions

$$E[Y_{i1}^{k \to j} \mid D_{ik0} = D_{ij1} = 1, X_i] = E[Y_{i1}^{j \to j} \mid D_{ik0} = D_{ij1} = 1, X_i], \tag{24}$$

$$E[Y_{i1}^{k \to 0} \mid D_{ik0} = D_{i01} = 1, X_i] = E[Y_{i1}^{0 \to 0} \mid D_{ik0} = D_{i01} = 1, X_i], \tag{25}$$

$$\text{and} \quad E[Y_{i1}^{0 \to k} \mid D_{i00} = D_{ik1} = 1, X_i] = E[Y_{i1}^{k \to k} \mid D_{i00} = D_{ik1} = 1, X_i] \tag{26}$$

$$= E[Y_{i1}^{j \to k} \mid D_{ij0} = D_{ik1} = 1, X_i] \tag{27}$$

hold for all $k \neq j$, then equations (21)-(23) also hold for $t = 0$.

As with Proposition 2, each time-1 mover average treatment effect in the multiple treatment setting is identified under parallel trends and effect homogeneity by a weighted average of outcome growth rates. Period-0 effects are also revealed under the stationarity assumptions (24)-(27). In fact, when $J = 2$ it can be verified that $\kappa_{i1t}^0 = \kappa_{i1t}$ and $\lambda_{i1t}^0 = \lambda_{i1t}$, so that equation (21) reduces to equation (12) and both the stationarity assumptions and the second and third support conditions of Propositions 4 and 2 coincide. Assumption CEH would furthermore be satisfied trivially, so these Propositions are equivalent in the binary treatment case.

When $J > 2$, the additional κ_{i1t}^k and λ_{i1t}^k terms use Assumption CEH to identify and weight together treatment effects for those moving between j and $k > 0$, along the lines of the above discussion. Figure 1(b) illustrates the additional mover and stayer groups that the weighting scheme implicitly contrasts, conditional on the controls, to identify these effects. For example, the time-1 treatment effect for those who move from treatment k to treatment j is revealed by the excess growth in outcomes between that group and those who move from treatment k to treatment 0 (the light solid lines in Figure 1(b)), while the same for those moving from j to k is identified by contrasting the growth moving from 0 to k and j to k, along with the growth of stayers in j and 0 (the dark solid lines). Contrasts within the groups differentiated by dashed and solid lines correspondingly reveal the additional time-0 treatment effects under the stationarity condition.

Although Proposition 4 suggests one estimation strategy for multi-valued treatment, there are in general infinitely-many weighting schemes that identify each MATE when Assumption CEH holds with multiple

treatments. This is because each of the contrasts of outcome growth within the groups shown in Figure 1(b) will produce the *same* average treatment effect for different k's, conditional on X_i. Therefore, provided these conditional effects are properly re-weighted to reflect the conditional distribution of other movers, each κ_{ijt}^k and λ_{ijt}^k pair, and thus all proper weighted averages across such pairs, will identify the MATEs. The following extension formalizes this intuition:

Corollary to Proposition 4: If $T = 2$ and the assumptions to Proposition 4 hold for either t, then

$$MATE_{jt} = E\left[\Delta Y_i \sum_{k \neq j} \left(W_\kappa^k \widetilde{\kappa}_{ijt}^k + W_\lambda^k \widetilde{\lambda}_{ijt}^k \right)\right], \tag{28}$$

where

$$\widetilde{\kappa}_{ijt}^k = \kappa_{ijt}^k \frac{E[D_{ijt}(1 - D_{ij,1-t}) \mid X_i]}{E[D_{ijt} D_{ik,1-t} \mid X_i]} \tag{29}$$

$$\text{and} \quad \widetilde{\lambda}_{ijt}^k = \lambda_{ijt}^k \frac{E[D_{ij,1-t}(1 - D_{ijt}) \mid X_i]}{E[D_{ij,1-t} D_{ikt} \mid X_i]}, \tag{30}$$

with κ_{ijt}^k and λ_{ijt}^k as in Proposition 4, and where $\sum_{k \neq j} W_\kappa^k = \sum_{k \neq j} W_\lambda^k = 1$.

Note that here Proposition 4 is a special case, where $W_\kappa^k = E[D_{ijt} D_{ik,1-t} \mid X_i]/E[D_{ijt}(1 - D_{ij,1-t}) \mid X_i]$ and $W_\lambda^k = E[D_{ij,1-t} D_{ikt} \mid X_i]/E[D_{ij,1-t}(1 - D_{ijt}) \mid X_i]$.

This corollary has two important implications for estimating MATEs in multi-valued treatment settings. First, it suggests an omnibus specification test: when the assumptions of Proposition 4 hold we should have, for each j and t and any two pairs of weight vectors (W_κ, W_λ) and (V_κ, V_λ),

$$H_0: \quad E\left[\Delta Y_i \left(\sum_{k \neq j} \left((W_\kappa^k - V_\kappa^k) \widetilde{\kappa}_{ijt}^k + (W_\lambda^k - V_\lambda^k) \widetilde{\lambda}_{ijt}^k \right)\right)\right] = 0, \tag{31}$$

which is a testable null. Second, given a set of first-step estimates of the various observable moments $E[D_{ijt} D_{ik,1-t} \mid X_i]$ that enter the $\widetilde{\kappa}_{ijt}^k$ and $\widetilde{\lambda}_{ijt}^k$, we can control the relative efficiency of two-step $MATE_{jt}$ estimators by selecting different weighting schemes. Namely, with Ω_{jt} denoting the asymptotic variance-covariance matrix of the vector $\widehat{\mu}_{jt}$ that collects the estimates $\widehat{E}[\Delta Y \widehat{\widetilde{\kappa}}_{ijt}^k]$ and $\widehat{E}[\Delta Y \widehat{\widetilde{\lambda}}_{ijt}^k]$ for all $k \neq j$ and $W = (W_\kappa', W_\lambda')'$ collecting the corresponding weights, we have by the above result that $W' \widehat{\mu}_{jt} \xrightarrow{p} MATE_{jt}$ and that

$$AVar(W' \widehat{\mu}_{jt}) = W' \Omega_{jt} W, \tag{32}$$

where $AVar(\cdot)$ denotes an estimator's asymptotic variance. Uniquely decomposing $\Omega_{jt} = \Pi_{jt}' \Psi_{jt} \Pi_{jt}$, where Π_{jt} is an upper-triangular matrix with diagonal elements equal to one and where Ψ_{jt} is a diagonal matrix, and labeling the first and second $J-1 \times J-1$ diagonal submatrices of Ψ_{jt}^{-1} as $\Psi_{\kappa jt}^{-1}$ and $\Psi_{\lambda jt}^{-1}$, respectively, the weighting vector

$$W^* = \Pi_{jt}^{-1} \left(\frac{diag(\Psi_{\kappa jt}^{-1})'}{trace(\Psi_{\kappa jt}^{-1})}, \frac{diag(\Psi_{\lambda jt}^{-1})'}{trace(\Psi_{\lambda jt}^{-1})} \right)' \tag{33}$$

will produce the efficient $MATE_{jt}$ estimator in this class, where $diag(\cdot)$ extracts the diagonal vector from a matrix and $trace(\cdot)$ sums its diagonal elements.[10] A sample analogue of this estimator can be readily formed given a consistent estimate of Ω_{jt}. Note that although in practice many propensity scores $E[D_{ij0}D_{ik1} \mid X_i]$ need to be approximated to form these estimators, this first step can in principle be distributed over multiple resources to reduce computational time, while calculation of the second weighting step should be near-instantaneous given the weights.

3 Extensions

Proposition 4 and its corollary provide a general approach to MATE estimation in two-period settings. I next consider two extensions to this framework which may prove useful for mover regressions in practice. First, I develop an estimation strategy in settings where a researcher wishes to compare average causal effects of different treatments across different mover subpopulations. Second, I extend all results to causal inference in multi-period mover designs.

3.1 Mutually-comparable Effects

In many mover designs, the omitted treatment category 0 is arbitrarily chosen. When effects are constant under Assumption CE, this choice does not limit the ability to causally compare any two treatments j and k, as we may estimate $E[Y_{it}^j - Y_{it}^k]$ by $\widehat{\beta}_j - \widehat{\beta}_k = \widehat{E}[Y_{it}^j - Y_{it}^0] - \widehat{E}[Y_{it}^k - Y_{it}^0]$, for any t. When effects are heterogeneous, however, comparability across MATEs is not guaranteed, as

$$\begin{aligned} &E[Y_{it}^j - Y_{it}^0 \mid \Delta D_{ij} \neq 0] - E[Y_{it}^k - Y_{it}^0 \mid \Delta D_{ik} \neq 0] \\ &= E[Y_{it}^j - Y_{it}^k \mid \Delta D_{ij} \neq 0] + (E[Y_{it}^k - Y_{it}^0 \mid \Delta D_{ij} \neq 0] - E[Y_{it}^k - Y_{it}^0 \mid \Delta D_{ik} \neq 0]) \\ &= E[Y_{it}^j - Y_{it}^k \mid \Delta D_{ik} \neq 0] + (E[Y_{it}^j - Y_{it}^0 \mid \Delta D_{ij} \neq 0] - E[Y_{it}^j - Y_{it}^0 \mid \Delta D_{ik} \neq 0]). \end{aligned} \tag{34}$$

Although the first terms in each of the second and third line have a causal interpretation, the second terms do not and are not necessarily equal to zero when Assumption CE fails to hold.

In cases where such comparability is needed, researchers may adopt a somewhat stronger restriction on treatment effect heterogeneity that is nevertheless still weaker than constant effects. To illustrate this we will need additional notation. Let the vector ΔD_i collect all ΔD_{ij}, and note that $\Delta D_i \neq 0$ if individual i moves move between any two treatments. Correspondingly, define the time-t *comparable mover average*

[10] That is, W^* solves $\min_W W'\Omega_{jt}W$ s.t. $\iota' W_\kappa^* = \iota' W_\lambda^* = 1$, where ι is a vector of ones. For example, if each $\widehat{\mu}_{jkt}$ estimator were mutually asymptotically independent (i.e. if U_{jt} were the identity matrix), the optimal linear combination would use inverse-variance weighting vectors W_κ^* and W_λ^* with respective elements $AVar(\widehat{E}[\Delta Y \widehat{\widetilde{\kappa}}_{ijt}^k])^{-1} / \sum_{\ell \neq j} AVar(\widehat{E}[\Delta Y \widehat{\widetilde{\kappa}}_{i\ell t}^k])^{-1}$ and $AVar(\widehat{E}[\Delta Y \widehat{\widetilde{\lambda}}_{ijt}^k])^{-1} / \sum_{\ell \neq j} AVar(\widehat{E}[\Delta Y \widehat{\widetilde{\lambda}}_{i\ell t}^k])^{-1}$. The Cholesky decomposition producing Π_{jt} and Ψ_{jt} generalizes this to the case of correlated estimators $\widehat{\mu}_{jkt}$.

treatment effect as

$$CMATE_{jt} = E[Y_{it}^{j} - Y_{it}^{0} \mid \Delta D_i \neq 0]. \tag{35}$$

Note that unlike with MATEs, these causal estimands are comparable across treatments: that is, for any t, j, and $k \neq j$, we have $E[Y_{it}^{j} - Y_{it}^{k} \mid \Delta D_i \neq 0] = CMATE_{jt} - CMATE_{kt}$

Next, let C_j denote the set of all variable-length n-tuples $C_j^p = (c_{j0}^p, \ldots, c_{jM(p)}^p)$, where $0 < M(p) < J$ and where $c_{jm}^p \in \{0, \ldots, J-1\}$, $c_{j0}^p = 0$, $c_{jM(p)}^p = j$, and $c_{jm}^p \neq c_{jn}^p$ for $m \neq n$. In words, C_j collects all possible *paths* C_j^p of length $M(p)$ from treatment 0 to treatment j via other treatment states c_{jm}^p, where no state is included in the path more than once.

Finally, consider the following restriction on effect heterogeneity:

Assumption CPI (*Conditional path ignorability*): For time period t, treatment j, and path C_j^p, the marginal effects $E[Y_{it}^{c_{jm}^p} - Y_{it}^{c_{j,m-1}^p} \mid \Delta D_i = d, X_i]$ do not depend on $d \neq 0$, for all m, provided these are well-defined

Assumption CPI may be thought of as further constraining the variability in potential outcomes, relative to Assumption CEH. When it holds, marginal mover effects along each step of the path C_j^p are mean-independent of the type of a mover considered, again conditional on the controls. We then have the following result:

Proposition 5: If $T = 2$ and, for some j and $t = 1$,

1. Assumption CPT holds for 0 and j, and Assumption CPI holds for t and some p
2. $P(D_{ic_{m-1}^p 0} = D_{ic_{m-1}^p 1} = 1 \mid X_i) > 0$, for all such p, all $0 < m \leq M(p)$, and all values of X_i for which $P(\Delta D_i \neq 0 \mid X_i)$,

then

$$CMATE_{jt} = E\left[\Delta Y_i \sum_p W_\rho^p \widetilde{\rho}_{ijt}^p\right], \tag{36}$$

where

$$\widetilde{\rho}_{ijt}^p = \frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)} \sum_{m=1}^{M(p)} \rho_{ijt}^p \tag{37}$$

for

$$\rho_{ijt}^p = (-1)^t D_{ic_{j,m-1}^p 1-t} \frac{D_{ic_{j,m-1}^p t} E[D_{ic_{j,m-1}^p 1-t} D_{ic_{jm}^p t} \mid X_i] - D_{ic_{jm}^p t} E[D_{ic_{j,m-1}^p 1-t} D_{ic_{j,m-1}^p t} \mid X_i]}{E[D_{ic_{j,m-1}^p 1-t} D_{ic_{j,m-1}^p t} \mid X_i] E[D_{ic_{j,m-1}^p 1-t} D_{ic_{jm}^p t} \mid X_i]}, \tag{38}$$

and where $\sum_p W_\rho^p = 1$. If the above conditions also hold for $t = 0$ and, moreover, the stationarity condition

$$E[Y_{i1}^{c_{jm}^p \to c_{j,m-1}^p} \mid D_{ic_{jm}^p 0} = D_{ic_{j,m-1}^p 1} = 1, X_i] = E[Y_{i1}^{c_{j,m-1}^p \to c_{j,m-1}^p} \mid D_{ic_{jm}^p 0} = D_{ic_{j,m-1}^p 1} = 1, X_i] \tag{39}$$

holds for all m in such p, then equations (36)-(38) also hold for $t = 0$.

This result shows that, under conditional parallel trends and the stronger restriction on treatment effects, comparable mover average treatment effects are identified by a weighting scheme involving some linear combination of the $\widetilde{\rho}^p_{ijt}$. Each of these combine causal effects for movers at each point in a given path C^p_j from treatment 0 to treatment j by leveraging Assumption CPI. As in Section 2.2, we can form specification tests and efficient estimators from multiple paths p, where the latter uses weights analogous to equation (33). Computation is again straightforward given first-step estimates of the propensity scores $E[D_{ij0}D_{ik1} \mid X_i]$.

3.2 Many Time Periods

Although we have thus far only considered comparisons between two time periods, it is straightforward to generalize the above results for MATE and CMATE identification in longer panels. Intuitively, in a multi-period setting we can imagine a modified version of Assumption CPT and either Assumption CEH or CPI holding for t and all $s < t$; we could then construct and average the two-period estimators of backward-looking MATEs and CMATEs over all such s. If we further assume the relevant stationarity assumptions hold we could additionally average in future periods $s > t$. Note that in these thought exercises, individuals deemed movers in one (s, t) pair may be considered stayers in another; a natural way to accommodate this in our definition of the target parameters is to define ΔD_{ij} to be a vector with elements $D_{ijt} - D_{ijs}$ for all t and $s \neq t$, and collect all such elements across j in ΔD_i. The movers represented by each $MATE_{jt}$ and $CMATE_{jt}$ are thus those who change treatment at any point in the sample. If we only wished to identify time-t causal effects for individuals who move up to time t (and thus not impose additional stationarity assumptions), we could instead define

$$\widetilde{MATE}_{jt} = E[Y^j_{it} - Y^0_{it} \mid \Delta D_{ijt} \neq 0] \tag{40}$$

$$\text{and} \quad \widetilde{CMATE}_{jt} = E[Y^j_{it} - Y^0_{it} \mid \Delta D_{it} \neq 0], \tag{41}$$

where ΔD_{ijt} and ΔD_{it} only collect $D_{ijt} - D_{ijs}$ for $s < t$.

Following the above intuition, consider the following generalization of conditional parallel trends,

Assumption CPT′: For time periods t and $s \neq t$ and treatment j, $E[Y^j_{it} - Y^j_{is} \mid D_{ijs} = D_{ijt} = 1, X_i] = E[Y^j_{it} - Y^j_{is} \mid D_{ijs} = D_{i0t} = 1, X_i] = E[Y^j_{it} - Y^j_{is} \mid D_{i0s} = D_{ijt} = 1, X_i]$, provided these are well-defined,

and, defining $\Delta_{ts}V_{it} = V_{it} - V_{is}$ as the difference in variable V_{it} between time periods s and t, a generalization of the conditional effect heterogeneity restriction,

Assumption CEH′: For time periods t and $s \neq t$ and treatment j, $E[Y^j_{it} - Y^0_{it} \mid \Delta_{ts}D_{ij} = d, \Delta_{ts}D_{ik} \neq 0, X_i]$ does not depend on $k \neq j$, for all $d \neq 0$, provided it is well-defined.

We then have a general $MATE_{jt}$ identification result,

Proposition 4′: If, for some j and some t,

1. Assumption CPT′ holds for t and all $s < t$, for treatments 0 and j, and Assumption CEH′ holds for t and all $s < t$ and $k \neq j$
2. $P(D_{i0t} = D_{iks} = 1 \mid X_i) > 0$ and $P(D_{ikt} = D_{iks} = 1 \mid X_i) > 0$ for all $s < t$ and k and all values of X_i for which $P(D_{ijt} = D_{iks} = 1 \mid X_i) > 0$
3. $P(D_{ikt} = D_{i0s} = 1 \mid X_i) > 0$, $P(D_{ijt} = D_{ijs} = 1 \mid X_i) > 0$, and $P(D_{i0t} = D_{i0s} = 1 \mid X_i) > 0$ for all $s < t$ and k and all values of X_i for which $P(D_{ikt} = D_{ijs} = 1 \mid X_i) > 0$,

then

$$\widetilde{MATE}_{jt} = \sum_{s<t} E\left[\Delta_{ts} Y_i \sum_{k\neq j} \left(W_\kappa^{ks} \widetilde{\kappa}_{ijt}^{ks} + W_\lambda^{ks} \widetilde{\lambda}_{ijt}^{ks}\right)\right], \tag{42}$$

where

$$\widetilde{\kappa}_{ijt}^{ks} = (-1)^{\mathbf{1}[t>s]} D_{iks} \frac{D_{i0t} E[D_{ijt} D_{iks} \mid X_i] - D_{ijt} E[D_{i0t} D_{iks} \mid X_i]}{E[D_{i0t} D_{iks} \mid X_i] E[D_{ijt} D_{iks} \mid X_i]} \frac{E[D_{ijt}(1 - D_{ijs}) \mid X_i]}{P(\Delta D_{ij} \neq 0)} \tag{43}$$

and

$$\begin{aligned} \widetilde{\lambda}_{ijt}^{ks} =& (-1)^{\mathbf{1}[t>s]} D_{ijs} \frac{D_{ikt} E[D_{ijt} D_{ijs} \mid X_i] - D_{ijt} E[D_{ikt} D_{ijs} \mid X_i]}{E[D_{ijt} D_{ijs} \mid X_i] E[D_{ikt} D_{ijs} \mid X_i]} \frac{E[D_{ijs}(1 - D_{ijt}) \mid X_i]}{P(\Delta D_{ij} \neq 0)} \\ &+ (-1)^{\mathbf{1}[t>s]} D_{i0s} \frac{D_{i0t} E[D_{ikt} D_{i0s} \mid X_i] - D_{ikt} E[D_{i0t} D_{i0s} \mid X_i]}{E[D_{ikt} D_{i0s} \mid X_i] E[D_{i0t} D_{i0s} \mid X_i]} \frac{E[D_{ijs}(1 - D_{ijt}) \mid X_i]}{P(\Delta D_{ij} \neq 0)}, \end{aligned} \tag{44}$$

and where $\sum_s \sum_{k\neq j} W_\kappa^{ks} = \sum_s \sum_{k\neq j} W_\lambda^{ks} = 1$. If the assumptions also hold for all $s > t$ and the stationarity conditions in Proposition 4 hold with t replacing 0 and s replacing 1 then moreover

$$MATE_{jt} = \sum_{s\neq t} E\left[\Delta_{ts} Y_i \sum_{k\neq j} \left(W_\kappa^{ks} \widetilde{\kappa}_{ijt}^{ks} + W_\lambda^{ks} \widetilde{\lambda}_{ijt}^{ks}\right)\right], \tag{45}$$

We similarly have an generalization of the result for comparable mover average treatment effects:

Proposition 5′: If, for some j and some t,

1. Assumption CPT′ holds for t and all $s < t$, for treatments 0 and j, and Assumption CPI holds for t and all p
2. $P(D_{ic_{m-1}^p t} = D_{ic_{m-1}^p s} = 1 \mid X_i) > 0$, for all $s < t$, p, and $0 < m \leq M(p)$, for all values of X_i for which $P(\Delta D_i \neq 0 \mid X_i) > 0$,

then

$$\widetilde{CMATE}_{jt} = \sum_s E\left[\Delta_{ts} Y_i \sum_p W_\rho^{ps} \widetilde{\rho}_{ijt}^{ps}\right], \tag{46}$$

where

$$\widetilde{\rho}_{ijt}^{ps} = \frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)} \sum_{m=1}^{M(p)} \rho_{ijt}^{ps} \tag{47}$$

for

$$\rho_{ijt}^{ps} = (-1)^{\mathbf{1}[t>s]} D_{ic_{j,m-1}^p s} \frac{D_{ic_{j,m-1}^p t} E[D_{ic_{j,m-1}^p s} D_{ic_{jm}^p t} \mid X_i] - D_{ic_{jm}^p t} E[D_{ic_{j,m-1}^p s} D_{ic_{j,m-1}^p t} \mid X_i]}{E[D_{ic_{j,m-1}^p s} D_{ic_{j,m-1}^p t} \mid X_i] E[D_{ic_{j,m-1}^p s} D_{ic_{jm}^p t} \mid X_i]}, \tag{48}$$

and where $\sum_s \sum_p W_\rho^{ps} = 1$. If the assumptions also hold for all $s > t$ and the stationarity conditions in Proposition 5 hold with t replacing 0 and s replacing 1 then moreover

$$CMATE_{jt} = \sum_s E\left[\Delta_{ts} Y_i \sum_p W_\rho^{ps} \widetilde{\rho}_{ijt}^{ps}\right], \tag{49}$$

Thus with multiple time periods satisfying the key assumptions, we may form estimators of each $MATE_{jt}$ and $CMATE_{jt}$ that aggregate comparisons across different outcome growth rates. Analogous results for the specification tests and efficient estimators derived in Section 2.2 would again apply to this most general case.

4 Conclusions

Although retaining the flavor of simpler difference-in-differences designs, mover experiments tend to require additional restrictions. Although causally interpretable in the binary treatment case, ordinary least squares regression estimates with many treatment alternatives fail to recover proper weighted averages of heterogeneous treatment effects under parallel trends alone. In contrast, the weighting estimators developed here accommodate limited effect variability that is not correlated with movement decisions conditional on controls. Whether such a restriction holds may be tested by overidentifying restrictions, and I show how to efficiently combine all mover information in such scenarios.

This two-step estimation procedure I propose requires the computation of approximations to a variety of different propensity scores; in principle this can be accomplished efficiently by a distributed (i.e. parallelized) first step. The second step requires only the computation of a particular set of weighted means given these approximations. The computational burden of estimation may thus prove light in practice, even relative to recent advances in traditional two-way fixed effect regressions (Abowd et al., 2002; Guimaraes and Portugal, 2010; Gaure, 2013; Correia, 2016). Future versions of this draft will consider these sorts of implementation details, as well as demonstrate the empirical relevance of the theory in prominent recent examples of mover studies.

As always, whether the assumptions outlined here are more plausible than other kinds of approaches to quasi-experimental mover estimation is ultimately a matter of context. Conditional parallel trends and effect homogeneity (or path ignorability) may be restrictions with the advantage of being both familiar to applied researchers and relatively straightforward consider in practice. At minimum, this approach allows researchers to easily verify the robustness of substantive conclusions drawn from conventional mover regressions

References

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

Abowd, J., R. Creecy, and F. Kramarz (2002): "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data," Cornell University Department of Economics Unpublished Working Paper.

Abowd, J. M., F. Kramarz, and D. Margolis (1999): "High Wage Workers and High Wage Firms," *Econometrica*, 67, 251–333.

Abowd, J. M., K. L. McKinney, and I. M. Schmutte (2015): "Modeling Endogenous Mobility in Wage Determination," Working Paper.

Allcott, H., R. Diamond, and J.-P. Dubé (2017): "The Geography of Poverty and Nutrition: Food Deserts and Food Choices Across the United States," Working Paper.

Andrews, D. W. K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 59, 307–345.

Angrist, J. and I. Fernandez-Val (2013): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress*, 3, 401–433.

Angrist, J. and M. Rokkanen (2015): "Wanna Get Away? Regression DIscontinuity Estimation of Exam School Effects Away from the Cutoff," *Journal of the American Statistical Association*, 110, 1331–1344.

Angrist, J. D. (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66, 249–288.

Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press.

Ashenfelter, O. (1978): "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47–57.

Ashenfelter, O. and D. Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648–660.

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

Bloom, N., J. Song, D. Price, F. Guvenen, and T. von Wachter (2015): "Firming up Inequality," *NBER Working Paper. 21199.*

Bonhomme, S., T. Lamadon, and E. Manresa (2017): "A Distributional Framework for Matched Employer Employee Data," .

Bronnenberg, B. J., J.-P. Dubé, and M. Gentzkow (2012): "The Evolution of Brand Preferences: Evidence from Consumer Migration," *American Economic Review*, 102, 2472–2508.

Card, D., J. Heining, and P. Kline (2013): "Workplace Heterogeneity and the Rise of West German Wage Inequality," *Quarterly Journal of Economics*, 128, 967–1015.

Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, 225–238.

——— (1982): "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, 5–46.

——— (1984): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator, Elsevier, vol. 2, chap. 22, 1247–1318, 1 ed.

Chernozhukov, V., I. Fernandez-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81, 535–580.

Chetty, R., J. Friedman, and J. Rockoff (2014): "Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104(9), 2593–2632.

Chetty, R. and N. Hendren (2017): "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects," *NBER Working Paper No. 23001*.

Correia (2016): "A Feasible Estimator for Linear Models with Multi-way Fixed Effects," Working Paper.

de Chaisemartin, C. and X. D'Haultfoeuille (2016): "Double Fixed Effects Estimators with Heterogeneous Treatment Effects," *Working Paper*.

Finkelstein, A., M. Gentzkow, P. Hull, and H. Williams (2017): "Adjusting Risk Adjustment: Accounting for Variation in Diagnostic Intensity," *New England Journal of Medicine*, 376, 608–610.

Finkelstein, A., M. Gentzkow, and H. Williams (2016): "Sources of Geographic Variation in Health Care: Evidence from Patient Migration," *Quarterly Journal of Economics*, 131, 1681–1726.

Gaure, S. (2013): "OLS with Multiple High Dimensional Category Variables," *Computational Statistics and Data Analysis*, 66, 8–18.

Graham, B. S. and J. L. Powell (2012): "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models," *Econometrica*, 80, 2105–2152.

Guimaraes, P. and P. Portugal (2010): "A Simple Feasible Procedure to Fit Models with High-Dimensional Fixed Effects," *Stata Journal*, 10, 628–649.

Hagedorn, M., T. H. Law, and I. Manovskii (2017): "Identifying Equilibrium Models of Labor Market Sorting," *Econometrica*, 85, 29–65.

Hahn, J. (2001): "Comment: Binary Regressors in Nonlinear Panel-Data Models with Fixed Effects," *Journal of Business and Economic Statistics*, 19, 16–17.

Heckman, J. J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Expermental Data," *Econometrica*, 66, 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching as an Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

Honore, B. E. (1992): "Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models wth Fixed Effects," *Econometrica*, 60, 533–565.

Hull, P. (2015): "IsoLATEing: Identifying Counterfactual-Specific Treatment Effects with Cross-Stratum Comparisons," Working Paper.

Imai, K. and I. S. Kim (2016): "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" Working Paper.

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

Jackson, C. K. (2013): "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers," *The Review of Economics and Statistics*, 95, 1096–1116.

Manski, C. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Response Data," *Econometrica*, 55, 357–362.

Robins, J. (1986): "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period: Application to Control of the Healthy Worker Survivor Effect," *Mathematical Modelling*, 7, 1393–1512.

——— (1997): "Causal Inference from Complex Longitudinal Data," in *Latent Variable Modeling and Applications to Causality: Lecture Notes in Statistics*, ed. by M. Berkane, Springer Verlag, vol. 120, 69–117.

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.

Sacarny, A. (2016): "Technological Diffusion Across Hospitals: The Case of a Revenue-Generating Practice," Working Paper.

Wooldridge, J. M. (2005): "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *The Review of Economics and Statistics*, 87, 385–390.

Yitzhaki, S. (1996): "On Using Linear Regressions in Welfare Economics," *Journal of Business and Economic Statistics*, 14, 478–486.

Appendix

Proof of Lemma 1

With $T = J = 2$ and $X_{it} = 0$, the mover regression can be written

$$\begin{aligned}\Delta Y_i &= \tau + \beta_1 \Delta D_{i1} + \Delta\epsilon_i \\ &= \tau + \beta_1(\mathbf{1}[\Delta D_{i1} = 1] - \mathbf{1}[\Delta D_{i1} = -1]) + \Delta\epsilon_i, \end{aligned} \tag{50}$$

which is nested by the regression

$$\begin{aligned}\Delta Y_i &= \widetilde{\tau} + \widetilde{\beta}_1(\mathbf{1}[\Delta D_{i1} = 1] - \mathbf{1}[\Delta D_{i1} = -1]) + \widetilde{\alpha}_1 \mathbf{1}[\Delta D_{i1} = -1] + \Delta\widetilde{\epsilon}_i \\ &= \widetilde{\tau} + \widetilde{\beta}_1 \mathbf{1}[\Delta D_{i1} = 1] + (\widetilde{\alpha}_1 - \widetilde{\beta}_1)\mathbf{1}[\Delta D_{it} = -1] + \Delta\widetilde{\epsilon}_i. \end{aligned} \tag{51}$$

This is a saturated model for $E[\Delta Y_i \mid \Delta D_{i1}]$, with

$$\widetilde{\beta}_1 = E[\Delta Y_i \mid \Delta D_{i1} = 1] - E[\Delta Y_i \mid \Delta D_{i1} = 0] \tag{52}$$

$$\text{and} \quad \widetilde{\alpha}_1 - \widetilde{\beta}_1 = E[\Delta Y_i \mid \Delta D_{i1} = -1] - E[\Delta Y_i \mid \Delta D_{i1} = 0]. \tag{53}$$

Thus, by usual omitted-variables bias logic,

$$\beta_1 = \widetilde{\beta}_1 + \widetilde{\alpha}_1 \frac{Cov(\mathbf{1}[\Delta D_{i1} = -1]), \Delta D_{i1})}{Var(\Delta D_{i1})}. \tag{54}$$

Define

$$\begin{aligned}\omega &= 1 + \frac{Cov(\mathbf{1}[\Delta D_{i1} = -1]), \Delta D_{i1})}{Var(\Delta D_{i1})} \\ &= \frac{Var(\mathbf{1}[\Delta D_{i1} = 1]) - Cov(\mathbf{1}[\Delta D_{i1} = 1], \mathbf{1}[\Delta D_{i1} = -1])}{Var(\mathbf{1}[\Delta D_{i1} = 1]) + Var(\mathbf{1}[\Delta D_{i1} = -1]) - 2Cov(\mathbf{1}[\Delta D_{i1} = 1], \mathbf{1}[\Delta D_{i1} = -1])}, \end{aligned} \tag{55}$$

and note that $\omega \in [0, 1]$. Thus,

$$\beta_1 = \widetilde{\beta}_1 \omega + (\widetilde{\beta}_1 - \widetilde{\alpha}_1)(1 - \omega). \tag{56}$$

Combining equations (52), (53), and (56) completes the proof. □

Proof of Proposition 1

Under the assumptions,

$$\begin{aligned}
&E[\Delta Y_i \mid \Delta D_{i1} = 1] - E[\Delta Y_i \mid \Delta D_{i1} = 0] \\
&= \left(E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 1] - E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 0, D_{i00} = 1]\right) p \\
&\quad + \left(E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 1] - E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 0, D_{i10} = 1]\right) (1-p) \\
&= \left(E[Y_{i1}^{0\to 1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1] - E[Y_{i1}^{0\to 0} - Y_{i0}^{0} \mid \Delta D_{i1} = 0, D_{i00} = 1]\right) p \\
&\quad + \left(E[Y_{i1}^{0\to 1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1] - E[Y_{i1}^{1\to 1} - Y_{i0}^{1} \mid \Delta D_{i1} = 0, D_{i10} = 1]\right) (1-p) \\
&= E[Y_{i1}^{1} - Y_{i1}^{0} \mid \Delta D_{i1} = 1]p + E[Y_{i0}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1](1-p), \qquad (57)
\end{aligned}$$

where $p = P(D_{i00} = 1 \mid \Delta D_{i1} = 0)$. Here the second equality follows from the definition of potential outcomes, while the third equality follows from the stationarity and parallel trends assumptions. Similarly,

$$\begin{aligned}
&E[\Delta Y_i \mid \Delta D_{i1} = 0] - E[\Delta Y_i \mid \Delta D_{i1} = -1] \\
&= \left(E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 0, D_{i00} = 1] - E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = -1]\right) p \\
&\quad + \left(E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 0, D_{i10} = 1] - E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = -1]\right) (1-p) \\
&= \left(E[Y_{i1}^{0\to 0} - Y_{i0}^{0} \mid \Delta D_{i1} = 0, D_{i00} = 1] - E[Y_{i1}^{1\to 0} - Y_{i0}^{1} \mid \Delta D_{i1} = -1]\right) p \\
&\quad + \left(E[Y_{i1}^{1\to 1} - Y_{i0}^{1} \mid \Delta D_{i1} = 0, D_{i10} = 1] - E[Y_{i1}^{1\to 0} - Y_{i0}^{1} \mid \Delta D_{i1} = -1]\right) (1-p) \\
&= E[Y_{i0}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = -1]p + E[Y_{i1}^{1} - Y_{i1}^{0} \mid \Delta D_{i1} = -1](1-p). \qquad (58)
\end{aligned}$$

Substituting these expressions in to the equation for the regression coefficient in Lemma 1 gives

$$\begin{aligned}
\beta_1 =& E[Y_{i1}^{1} - Y_{i1}^{0} \mid \Delta D_{i1} = 1]p\omega + E[Y_{i1}^{1} - Y_{i1}^{0} \mid \Delta D_{i1} = -1](1-p)(1-\omega) \qquad (59) \\
&+ E[Y_{i0}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1](1-p)\omega + E[Y_{i0}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = -1]p(1-\omega),
\end{aligned}$$

completing the proof. □

Proof of Proposition 2

Let $\psi_{ij} = P(\Delta D_{i1} \neq 0)/E[D_{ij0}D_{i01} \mid X_i]$ and $\omega_{ij} = P(\Delta D_{i1} \neq 0)/E[D_{i00}D_{ij1} \mid X_i]$. We have, for all values of X_i where $P(\Delta D_{ij} = -1 \mid X_i) > 0$,

$$\begin{aligned}
E[\Delta Y_i \kappa_{i10} \psi_{i1} \mid X_i] =& E\left[\Delta Y_i D_{i01} \frac{E[D_{i10}D_{i01} \mid X_i] - D_{i10}E[D_{i01} \mid X_i]}{E[D_{i00}D_{i01} \mid X_i]E[D_{i10}D_{i01} \mid X_i]} \mid X_i\right] \\
=& \frac{E[\Delta Y_i \mid D_{i01} = 1, X_i] - E[\Delta Y_i \mid D_{i01} = D_{i10} = 1, X_i]}{P(D_{i00} = D_{i01} = 1 \mid D_{i01} = 1, X_i)} \\
=& E[\Delta Y_i \mid D_{i00} = D_{i01} = 1, X_i] - E[\Delta Y_i \mid \Delta D_{i1} = -1, X_i] \\
=& E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} = -1, X_i] \qquad (60)
\end{aligned}$$

where the last equality follows by Assumption CPT and the conditional stationarity restriction. Similarly, for all values of the covariate vector for which $P(\Delta D_{ij} = 1 \mid X_i) > 0$,

$$E[\Delta Y_i \lambda_{i10} \omega_{i1} \mid X_i] = E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} = 1, X_i] \qquad (61)$$

The period-0 mover average treatment effect is thus

$$\begin{aligned}
E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} \neq 0] =& \int E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} \neq 0, X_i] dP(X_i \mid \Delta D_{i1} \neq 0) \\
=& \int E[\Delta Y_i \kappa_{i10} \psi_{i1} \mid X_i] \frac{P(\Delta D_{i1} = -1 \mid X_i)}{P(\Delta D_{i1} \neq 0 \mid X_i)} dP(X_i \mid \Delta D_{i1} \neq 0) \\
&+ \int E[\Delta Y_i \lambda_{i10} \omega_{i1} \mid X_i] \frac{P(\Delta D_{i1} = 1 \mid X_i)}{P(\Delta D_{i1} \neq 0 \mid X_i)} dP(X_i \mid \Delta D_{i1} \neq 0) \\
=& E\left[\Delta Y_i \kappa_{i10} \psi_{i1} \frac{P(\Delta D_{i1} = -1 \mid X_i)}{P(\Delta D_{i1} \neq 0)} + \Delta Y_i \lambda_{i10} \omega_{i1} \frac{P(\Delta D_{i1} = 1 \mid X_i)}{P(\Delta D_{i1} \neq 0)}\right] \\
=& E[\Delta Y_i (\kappa_{i10} + \lambda_{i10})]. \qquad (62)
\end{aligned}$$

The same steps prove that $E[Y_{i1}^1 - Y_{i1}^0 \mid \Delta D_{i1} \neq 0] = E[\Delta Y_i(\lambda_{i11} + \kappa_{i11})]$, since

$$E[\Delta Y_i \lambda_{i11} \psi_{i1} \mid X_i] = E[Y_{i1}^1 - Y_{i1}^0 \mid \Delta D_{i1} = -1, X_i] \qquad (63)$$

$$\text{and} \quad E[\Delta Y_i \kappa_{i11} \omega_{i1} \mid X_i] = E[Y_{i1}^1 - Y_{i1}^0 \mid \Delta D_{i1} = 1, X_i], \qquad (64)$$

for all values of X_i for which the right-hand side of these equations are well-defined. Note that in this case no conditional stationarity assumption is required. □

Proof of Corollary to Proposition 2

Note that, for each j and t,

$$\begin{aligned}
E\left[\Delta Y_i \frac{(\kappa_{ijt} + \lambda_{ijt})(M_{ij} + S_{ij})}{P(M_{ij} = 1 \mid \Delta D_{i1} \neq 0)}\right] &= E\left[\Delta Y_i (\kappa_{ijt} + \lambda_{ijt}) \mid M_{ij} + S_{ij} = 1\right] P(M_{ij} + S_{ij} = 1) \frac{P(\Delta D_{i1} \neq 0)}{P(M_{ij} = 1)} \\
&= E\left[\Delta Y_i \frac{(\kappa_{ijt} + \lambda_{ijt}) P(\Delta D_{i1} \neq 0)}{P(\Delta D_{i1} \neq 0 \mid M_{ij} + S_{ij} = 1)} \mid M_{ij} + S_{ij} = 1\right], \qquad (65)
\end{aligned}$$

The result then follows similarly as the proof to Proposition 2, provided $M_{ij} + S_{ij}$ is included in X_i. □

Proof of Proposition 3

I prove this for the case of $T = 2$ and $X_{it} = 0$, though similar steps prove the Proposition more generally. The mover regression is

$$\Delta Y_i = \tau + \sum_{j \neq 0} \beta_j (\mathbf{1}[\Delta D_{i1} = 1] - \mathbf{1}[\Delta D_{i1} = -1]) + \Delta \epsilon_i. \tag{66}$$

This specification is nested by the regression

$$\Delta Y_i = \widetilde{\tau} + \sum_{j \neq 0} \widetilde{\beta}_j (\mathbf{1}[\Delta D_{ij} = 1] - \mathbf{1}[\Delta D_{ij} = -1]) + \sum_k \sum_{j \neq 0,k} \widetilde{\delta}_{jk} \mathbf{1}[\Delta D_{ik} = 1] \mathbf{1}[\Delta D_{ij} = -1] + \Delta \widetilde{\epsilon}_i, \tag{67}$$

which can be written

$$\Delta Y_i = \widetilde{\tau} + \sum_{j \neq 0} \widetilde{\beta}_j D_{i00} D_{ij1} + \sum_{j \neq 0} (\widetilde{\delta}_{j0} - \widetilde{\beta}_j) D_{ij0} D_{i01} + \sum_k \sum_{j \neq 0,k} (\widetilde{\delta}_{jk} + \widetilde{\beta}_k - \widetilde{\beta}_j) D_{ij0} D_{ik1} + \Delta \widetilde{\epsilon}_i. \tag{68}$$

Equation (68) is a saturated model for $E[\Delta Y_i \mid \{D_{ij0} D_{ik1}\}_{k \neq j}]$, with

$$\widetilde{\beta}_j = E[\Delta Y_i \mid D_{i00} = D_{ij1} = 1] - E[\Delta Y_i \mid D_{i\ell 0} = D_{i\ell 1}, \forall \ell], \tag{69}$$

$$\widetilde{\delta}_{j0} - \widetilde{\beta}_j = E[\Delta Y_i \mid D_{ij0} = D_{i01} = 1] - E[\Delta Y_i \mid D_{i\ell 0} = D_{i\ell 1}, \forall \ell], \tag{70}$$

$$\text{and} \quad \widetilde{\delta}_{jk} + \widetilde{\beta}_k - \widetilde{\beta}_j = E[\Delta Y_i \mid D_{ij0} = D_{ik1} = 1] - E[\Delta Y_i \mid D_{i\ell 0} = D_{i\ell 1}, \forall \ell]. \tag{71}$$

Under Assumption CPT and stationarity, we can write

$$\begin{aligned}
\widetilde{\beta}_j =& (E[Y_{i1}^j - Y_{i0}^0 \mid D_{i00} = D_{ij1} = 1] - E[Y_{i1}^0 - Y_{i0}^0 \mid D_{i00} = D_{i01} = 1]) p_0 \\
&+ (E[Y_{i1}^j - Y_{i0}^0 \mid D_{i00} = D_{ij1} = 1] - E[Y_{i1}^j - Y_{i0}^j \mid D_{ij0} = D_{ij1} = 1]) p_j \\
&+ \sum_{\ell \neq 0,j} (E[Y_{i1}^j - Y_{i0}^0 \mid D_{i00} = D_{ij1} = 1] - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} = D_{i\ell 1} = 1]) p_\ell \\
=& E[Y_{i1}^j - Y_{i1}^0 \mid D_{i00} = D_{ij1} = 1] p_0 + E[Y_{i0}^j - Y_{i0}^0 \mid D_{i00} = D_{ij1} = 1](1 - p_0) \\
&+ \sum_{\ell \neq 0,j} (E[Y_{i1}^j - Y_{i0}^j \mid D_{ij0} = D_{ij1} = 1] - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} = D_{i\ell 1} = 1]) p_\ell,
\end{aligned} \tag{72}$$

where $p_k = P(D_{ik0} = 1 \mid D_{i\ell 0} = D_{i\ell 1}, \forall \ell)$. Similarly, we have

$$\begin{aligned}
\widetilde{\delta}_{j0} =& \widetilde{\beta}_j - (E[Y_{i1}^j - Y_{i1}^0 \mid D_{ij0} = D_{i01} = 1] p_0 + E[Y_{i0}^j - Y_{i0}^0 \mid D_{ij0} = D_{i01} = 1](1 - p_0)) \\
&- \sum_{\ell \neq 0,j} p_\ell (E[Y_{i1}^j - Y_{i0}^j \mid D_{ij0} = D_{ij1} = 1] - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} = D_{i\ell 1} = 1]) \\
=& (E[Y_{i1}^j - Y_{i0}^0 \mid D_{i00} = D_{ij1} = 1] - E[Y_{i1}^j - Y_{i1}^0 \mid D_{ij0} = D_{i01} = 1]) p_0 \\
&+ (E[Y_{i0}^j - Y_{i0}^0 \mid D_{i00} = D_{ij1} = 1] - E[Y_{i0}^j - Y_{i0}^0 \mid D_{ij0} = D_{i01} = 1])(1 - p_0))
\end{aligned} \tag{73}$$

and

$$
\begin{aligned}
\widetilde{\delta}_{jk} =& \widetilde{\beta}_j - \widetilde{\beta}_k + E[Y_{i1}^k - Y_{i1}^j \mid D_{ij0} = D_{ik1} = 1]p_j + E[Y_{i0}^k - Y_{i0}^j \mid D_{ij0} = D_{ik1} = 1](1-p_j) \\
&+ \sum_{\ell \neq j,k} (E[Y_{i1}^k - Y_{i0}^k \mid D_{ik0} = D_{ik1}, \forall \ell]) - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} = D_{i\ell 1}, \forall \ell])p_\ell \\
=& E[Y_{i1}^j - Y_{i1}^0 \mid D_{i00} = D_{ij1} = 1]p_0 + E[Y_{i0}^j - Y_{i0}^0 \mid D_{i00} = D_{ij1} = 1](1-p_0) \\
&+ \sum_{\ell \neq 0,j} (E[Y_{i1}^j - Y_{i0}^j \mid D_{ij0} = D_{ij1} = 1] - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} = D_{i\ell 1} = 1])p_\ell \\
&- E[Y_{i1}^k - Y_{i1}^0 \mid D_{i00} = D_{ik1} = 1]p_0 - E[Y_{i0}^k - Y_{i0}^0 \mid D_{i00} = D_{ik1} = 1](1-p_0) \\
&- \sum_{\ell \neq 0,k} (E[Y_{i1}^k - Y_{i0}^k \mid D_{ik0} = D_{ik1} = 1] - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} = D_{i\ell 1} = 1])p_\ell, \\
&+ E[Y_{i1}^k - Y_{i1}^j \mid D_{ij0} = D_{ik1} = 1]p_j + E[Y_{i0}^k - Y_{i0}^j \mid D_{ij0} = D_{ik1} = 1](1-p_j) \\
&+ \sum_{\ell \neq j,k} (E[Y_{i1}^k - Y_{i0}^k \mid D_{ik0} = D_{ik1}, \forall \ell]) - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} = D_{i\ell 1}, \forall \ell])p_\ell \\
=& (E[Y_{i1}^j - Y_{i1}^0 \mid D_{i00} = D_{ij1} = 1] + E[Y_{i0}^k - Y_{i0}^j \mid D_{ij0} = D_{ik1} = 1])p_0 \\
&- E[Y_{i1}^k - Y_{i1}^0 \mid D_{i00} = D_{ik1} = 1]p_0 \\
&+ (E[Y_{i0}^j - Y_{i0}^0 \mid D_{i00} = D_{ij1} = 1] + E[Y_{i1}^k - Y_{i1}^j \mid D_{ij0} = D_{ik1} = 1])p_j \\
&- E[Y_{i0}^k - Y_{i0}^0 \mid D_{i00} = D_{ik1} = 1])p_j \\
&+ (E[Y_{i0}^j - Y_{i0}^0 \mid D_{i00} = D_{ij1} = 1] + E[Y_{i0}^k - Y_{i0}^j \mid D_{ij0} = D_{ik1} = 1])(1-p_0-p_j) \\
&- E[Y_{i0}^k - Y_{i0}^0 \mid D_{i00} = D_{ik1} = 1](1-p_0-p_j) \\
&+ \sum_{\ell \neq 0,j} (E[Y_{i1}^j - Y_{i0}^j \mid D_{ij0} = D_{ij1} = 1] - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} = D_{i\ell 1} = 1])p_\ell \\
&- \sum_{\ell \neq 0,k} (E[Y_{i1}^k - Y_{i0}^k \mid D_{ik0} = D_{ik1} = 1] - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} = D_{i\ell 1} = 1])p_\ell, \\
&+ \sum_{\ell \neq j,k} (E[Y_{i1}^k - Y_{i0}^k \mid D_{ik0} = D_{ik1}, \forall \ell]) - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} = D_{i\ell 1}, \forall \ell])p_\ell
\end{aligned} \tag{74}
$$

Finally, note that we can the standard omitted-variables bias formula to write the vector of mover regression coefficients in terms of the saturated model's coefficient vector:

$$
\beta = \widetilde{\beta} + \sum_k \sum_{j \neq 0,k} \widetilde{\delta}_{jk} R_{jk}, \tag{75}
$$

where R_{jk} denotes the (generally non-zero) coefficient vector from regressing each $\mathbf{1}[\Delta D_{ik} = 1]\mathbf{1}[\Delta D_{ij} = -1]$ on the set of $\Delta D_{i\ell}$ for $\ell > 0$. Substituting equations (72)-(74) in to this expression shows that β will not in general identify a weighted average of causal parameters. $\square$.

Proof of Proposition 4

Let $\psi_{ij}^k = P(\Delta D_{ij} \neq 0)/E[D_{ij0}D_{ik1} \mid X_i]$ and $\omega_{ij}^k = P(\Delta D_{ij} \neq 0)/E[D_{ik0}D_{ij1} \mid X_i]$. As in the proof to Proposition 2, we have for all $k \neq j$ and all values of X_i such that $P(D_{ij0} = D_{ik1} = 1 \mid X_i) > 0$,

$$\begin{aligned}
E[\Delta Y_i \kappa_{ij0}^k \psi_{ij}^k \mid X_i] =& E\left[\Delta Y_i D_{ik1} \frac{D_{i00}E[D_{ij0}D_{ik1} \mid X_i] - D_{ij0}E[D_{i00}D_{ik1} \mid X_i]}{E[D_{i00}D_{ik1} \mid X_i]E[D_{ij0}D_{ik1} \mid X_i]} \mid X_i\right] \\
=& \frac{E[\Delta Y_i D_{i00}D_{ik1} \mid X_i]}{P(D_{i00} = D_{ik1} = 1 \mid X_i)} - \frac{E[\Delta Y_i D_{ij0}D_{ik1} \mid X_i]}{P(D_{ij0} = D_{ik1} = 1 \mid X_i)} \\
=& E[\Delta Y_i \mid D_{i00} = D_{ik1} = 1, X_i] - E[\Delta Y_i \mid D_{ij0} = D_{ik1} = 1, X_i] \\
=& E[Y_{i0}^k - Y_{i0}^0 \mid D_{i00} = D_{ik1} = 1, X_i] - E[Y_{i0}^k - Y_{i0}^j \mid D_{ij0} = D_{ik1} = 1, X_i] \\
=& E[Y_{i0}^j - Y_{i0}^0 \mid D_{ij0} = D_{ik1} = 1, X_i], \qquad (76)
\end{aligned}$$

where the fourth equality follows by Assumption CPT and the stationarity condition, after subtracting and adding $E[\Delta Y_i \mid D_{ik0} = D_{ik1} = 0]$, and the last line is due to Assumption CEH. Similarly, for all k and values of X_i such that $P(D_{ik0} = D_{ij1} = 1 \mid X_i) > 0$,

$$\begin{aligned}
E[\Delta Y_i \lambda_{ij0}^k \omega_{ij}^k \mid X_i] =& E\left[\Delta Y_i D_{ij1} \frac{D_{ik0}E[D_{ij0}D_{ij1} \mid X_i] - D_{ij0}E[D_{ik0}D_{ij1} \mid X_i]}{E[D_{ij0}D_{ij1} \mid X_i]E[D_{ik0}D_{ij1} \mid X_i]} \mid X_i\right] \\
&+ E\left[\Delta Y_i D_{i01} \frac{D_{i00}E[D_{ik0}D_{i01} \mid X_i] - D_{ik0}E[D_{i00}D_{i01} \mid X_i]}{E[D_{ik0}D_{i01} \mid X_i]E[D_{i00}D_{i01} \mid X_i]} \mid X_i\right] \\
=& E[\Delta Y_i \mid D_{ik0} = D_{ij1} = 1, X_i] - E[\Delta Y_i \mid D_{ij0} = D_{ij1} = 1, X_i] \\
&+ E[\Delta Y_i \mid D_{i00} = D_{i01} = 1, X_i] - E[\Delta Y_i \mid D_{ik0} = D_{i01} = 1, X_i] \\
=& E[Y_{i0}^j - Y_{i0}^k \mid D_{ik0} = D_{ij1} = 1, X_i] + E[Y_{i0}^k - Y_{i0}^0 \mid D_{ik0} = D_{i01} = 1, X_i] \\
=& E[Y_{i0}^j - Y_{i0}^0 \mid D_{ik0} = D_{ij1} = 1, X_i], \qquad (77)
\end{aligned}$$

again by Assumptions CPT and CEH and the stationarity condition. Thus,

$$\begin{aligned}
E[Y_{i0}^j - Y_{i0}^0 \mid \Delta D_{ij} \neq 0] =& \sum_{k \neq j} \int E[\Delta Y_i \kappa_{ij0}^k \psi_{ij}^k \mid X_i] \frac{P(D_{ij0} = D_{ik1} = 1 \mid X_i)}{P(\Delta D_{ij} \neq 0 \mid X_i)} dP(X_i \mid \Delta D_{ij} \neq 0) \\
&+ \sum_{k \neq j} \int E[\Delta Y_i \lambda_{ij0}^k \omega_{ij}^k \mid X_i] \frac{P(D_{ik0} = D_{ij1} = 1 = 1 \mid X_i)}{P(\Delta D_{ij} \neq 0 \mid X_i)} dP(X_i \mid \Delta D_{ij} \neq 0) \\
=& E\left[\Delta Y_i \sum_{k \neq j} (\kappa_{ij0}^k + \lambda_{ij0}^k)\right]. \qquad (78)
\end{aligned}$$

The same steps prove the $E[Y_{i1}^j - Y_{i1}^0 \mid \Delta D_{ij} \neq 0] = E\left[\Delta Y_i \sum_{k \neq j} (\kappa_{ij1}^k + \lambda_{ij1}^k)\right]$, since for all $k \neq j$,

$$\begin{aligned}
E[\Delta Y_i \lambda_{ij0}^k \psi_{ij}^k \mid X_i] =& E[Y_{i0}^j - Y_{i0}^0 \mid D_{ij0} = D_{ik1} = 1, X_i] \qquad (79) \\
\text{and} \quad E[\Delta Y_i \kappa_{ij0}^k \omega_{ij}^k \mid X_i] =& E[Y_{i0}^j - Y_{i0}^0 \mid D_{ik0} = D_{ij1} = 1, X_i], \qquad (80)
\end{aligned}$$

for all values of X_i for which the right-hand side of these equations are well-defined. As with Proposition 2, note that the stationarity condition is not required for these equalities. □

Proof of the Corollary to Propositions 4 and to Proposition 4′

As in the proof to Proposition 4, we have under Assumption CPT,

$$E\left[\Delta Y_i \frac{\kappa^k_{ij0} E[D_{ij0}(1-D_{ij1}) \mid X_i]}{E[D_{ij0}D_{ik1} \mid X_i]} \mid X_i\right] = E\left[\Delta Y_i \kappa^k_{ij0} \psi^k_{ij} \frac{P(\Delta D_{ij} = -1 \mid X_i)}{P(\Delta D_{ij} \neq 0)} \mid X_i\right]$$
$$= E[Y^j_{i0} - Y^0_{i0} \mid D_{ij0} = D_{ik1} = 1, X_i] \frac{P(\Delta D_{ij} = -1 \mid X_i)}{P(\Delta D_{ij} \neq 0)} \quad (81)$$

$$\text{and } E\left[\Delta Y_i \frac{\lambda^k_{ij0} E[D_{ij1}(1-D_{ij0}) \mid X_i]}{E[D_{ik0}D_{ij1} \mid X_i]} \mid X_i\right] = E[Y^j_{i0} - Y^0_{i0} \mid D_{ik0} = D_{ij1} = 1, X_i] \frac{P(\Delta D_{ij} = 1 \mid X_i)}{P(\Delta D_{ij} \neq 0)}, \quad (82)$$

While under Assumption CEH, $E[Y^j_{i0} - Y^0_{i0} \mid D_{ij0} = D_{ik1} = 1, X_i] = E[Y^j_{i0} - Y^0_{i0} \mid \Delta D_{ij} = -1, X_i]$ and $E[Y^j_{i0} - Y^0_{i0} \mid D_{ik0} = D_{ij1} = 1, X_i] = E[Y^j_{i0} - Y^0_{i0} \mid \Delta D_{ij} = 1]$ for all $k \neq j$. Thus for any set of W^k_κ and W^k_λ,

$$\begin{aligned} E[Y^j_{i0} - Y^0_{i0} \mid \Delta D_{ij} \neq 0] &= \int E[Y^j_{i0} - Y^0_{i0} \mid \Delta D_{ij} = -1, X_i] dP(X_i \mid \Delta D_{ij} = -1) \frac{P(\Delta D_{ij} = -1)}{P(\Delta D_{ij} \neq 0)} \\ &\quad + \int E[Y^j_{i0} - Y^0_{i0} \mid \Delta D_{ij} = 1, X_i] dP(X_i \mid \Delta D_{ij} = 1) \frac{P(\Delta D_{ij} = 1)}{P(\Delta D_{ij} \neq 0)} \\ &= \int \left(\sum_{k \neq j} W^k_\kappa E[\Delta Y_i \widetilde{\kappa}^k_{ij0} \mid X_i] \right) \frac{P(\Delta D_{ij} = -1)}{P(\Delta D_{ij} = -1 \mid X_i)} dP(X_i \mid \Delta D_{ij} = -1) \\ &\quad + \int \left(\sum_{k \neq j} W^k_\lambda E[\Delta Y_i \widetilde{\lambda}^k_{ijt} \mid X_i] \right) \frac{P(\Delta D_{ij} = 1)}{P(\Delta D_{ij} = 1 \mid X_i)} dP(X_i \mid \Delta D_{ij} = 1) \\ &= E\left[\Delta Y_i \sum_{k \neq j} (W^k_\kappa \widetilde{\kappa}^k_{ijt} + W^k_\lambda \widetilde{\lambda}^k_{ij0}) \right] \end{aligned} \quad (83)$$

The same steps prove the analogous result for $MATE_{j1}$, since

$$E\left[\Delta Y_i \frac{\kappa^k_{ij1} E[D_{ij1}(1-D_{ij0}) \mid X_i]}{E[D_{ik0}D_{ij1} \mid X_i]} \mid X_i\right] = E[Y^j_{i1} - Y^0_{i1} \mid D_{ik0} = D_{ij1} = 1, X_i] \frac{P(\Delta D_{ij} = 1 \mid X_i)}{P(\Delta D_{ij} \neq 0)} \quad (84)$$

$$\text{and } E\left[\Delta Y_i \frac{\lambda^k_{ij1} E[D_{ij0}(1-D_{ij1}) \mid X_i]}{E[D_{ij0}D_{ik1} \mid X_i)} \mid X_i\right] = E[Y^j_{i1} - Y^0_{i1} \mid D_{ij0} = D_{ik1} = 1, X_i] \frac{P(\Delta D_{ij} = -1 \mid X_i)}{P(\Delta D_{ij} \neq 0)}. \quad (85)$$

It is similarly straightforward to verify Proposition 4′ by this logic. □

Proof of Propositions 5 and 5′

Note that

$$\begin{aligned}
E[\Delta Y_i \rho^{\rho}_{ijt}] =& E[Y^{c^p_{j,m-1}}_{i,1-t} - Y^{c^p_{j,m-1}}_{it} \mid D_{ic^p_{j,m-1}1-t} = D_{ic^p_{j,m-1}t} = 1, X_i] \\
&- E[Y^{c^p_{j,m-1}}_{i,1-t} - Y^{c^p_{jm}}_{it} \mid D_{ic^p_{j,m-1}1-t} = D_{ic^p_{jm}t} = 1, X_i] \\
=& E[Y^{c^p_{jm}}_{it} - Y^{c^p_{j,m-1}}_{it} \mid D_{ic^p_{j,m-1}1-t} = D_{ic^p_{jm}t} = 1, X_i] \\
=& E[Y^{c^p_{jm}}_{it} - Y^{c^p_{j,m-1}}_{it} \mid \Delta D_i \neq 0, X_i], \qquad (86)
\end{aligned}$$

where the last two lines follow from Assumptions CPT (and the stationarity condition in the case of $t = 0$) and CPI, respectively. Thus, for any p,

$$E[\Delta Y_i \widetilde{\rho}^p_{ijt} \mid X_i] = E[Y^j_{it} - Y^0_{it} \mid \Delta D_i \neq 0, X_i] \frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)} \qquad (87)$$

and for any weights W^p_ρ with $\sum_p W^p_\rho = 1$,

$$\begin{aligned}
E[Y^j_{it} - Y^0_{it} \mid \Delta D_i \neq 0] &= \int E[Y^j_{it} - Y^0_{it} \mid \Delta D_i \neq 0, X_i] dP(X_i \mid \Delta D_i \neq 0) \\
&= \int E[\Delta Y_i \sum_k W^p_\rho \widetilde{\rho}^p_{ijt} \mid X_i] \frac{P(\Delta D_i \neq 0)}{P(\Delta D_i \neq 0 \mid X_i)} dP(X_i \mid \Delta D_i \neq 0) \qquad (88) \\
&= E[\Delta Y_i \sum_k W^p_\rho \widetilde{\rho}^p_{ijt}], \qquad (89)
\end{aligned}$$

completing the proof. Similar steps may be used to prove Proposition 5′. □