Estimating Treatment Effects in Mover Designs[*]

Peter Hull[†]

April 2018

Abstract

Researchers increasingly leverage movement across multiple treatments to estimate causal effects. While these "mover regressions" are often motivated by a linear constant-effects model, it is not clear what they capture under weaker quasi-experimental assumptions. I show that binary treatment mover regressions recover a convex average of four difference-in-difference comparisons and are thus causally interpretable under a standard parallel trends assumption. Estimates from multiple-treatment models, however, need not be causal without stronger restrictions on the heterogeneity of treatment effects and time-varying shocks. I propose a class of two-step estimators to isolate and combine the large set of difference-in-difference quasi-experiments generated by a mover design, identifying mover average treatment effects under conditional-on-covariate parallel trends and effect homogeneity restrictions. I characterize the efficient estimators in this class and derive specification tests based on the model's overidentifying restrictions. Future drafts will apply the theory to the Finkelstein et al. (2016) movers design, analyzing the causal effects of geography on healthcare utilization.

[*] I thank Sarah Abraham, Alberto Abadie, Isaiah Andrews, Josh Angrist, Amy Finkelstein, Matt Gentzkow, Paul Goldsmith-Pinkham, Sida Peng, Nathan Hendren, Liyang Sun, Chris Walters, and Heidi Williams for valuable feedback. This draft is a work in progress and should not be cited without prior notification.

[†] University of Chicago and Microsoft Research. Email: hull@uchicago.edu; website: http://peterhull.net

1 Introduction

The rise of rich longitudinal data has broadened the scope for causal inference in economics. Rather than estimating a single cross-sectional treatment effect, researchers increasingly exploit variation in an individual's treatment 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 a large number of causal effects simultaneously.[1] Often these effects are obtained from a linear two-way fixed effects regression, motivated by a static, additive, and constant-effect model (e.g. Abowd et al. (1999)). It is usually not clear, however, what these sorts of "mover regressions" capture under misspecifications of the model, including heterogeneous treatment effects, outcome persistence, and time-varying shocks.

This paper explores the causal content of mover designs in a treatment effects framework, relaxing the canonical regression model with familiar quasi-experimental restrictions. In the simplest case of a binary treatment, two time periods, and no additional controls, I show that a mover regression identifies a weighted average of four difference-in-difference comparisons, and is therefore causally interpretable under a restriction on outcome persistance and a standard parallel trends assumption. This result links mover analyses to simpler quasi-experimental designs, but does not easily extend to settings with multiple unordered treatments. I show that, in general, mover regressions need not identify weighted averages of heterogeneous causal effects under parallel trends alone; rather they require additional restrictions on potential outcome heterogeneity across both treatments and time.

Motivated by these results, I develop a class of two-step *mover average treatment effect* (MATE) estimators for quasi-experimental mover designs. The estimators can be thought to extend the semiparametric difference-in-difference approach of Abadie (2005) to settings where individuals move both into and out of multiple treatments over time. Identification follows from conditional-on-covariate restrictions on trends and treatment effect heterogeneity, which are satisfied by a partially-separable model of dynamic potential outcomes. Certain MATEs are identified without direct restrictions on potential outcome persistence. The key effect homogeneity assumption permits extrapolation across observably-similar individuals from the many difference-in-difference comparisons embeded in mover designs and generates a large set of overidentifying restrictions. I characterize efficient MATE estimators and omnibus specification tests of these restrictions, both of which are straightforward to compute given a set of first-step propensity score estimates.

[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), Finkelstein et al. (2017), Molitor (2017), Allcott et al. (2017), and Chetty and Hendren (Forthcoming).

This analysis contributes to a small but growing econometric literature relaxing the canonical assumptions of two-way fixed effect regressions, typically as applied to matched worker-firm panels. Abowd et al. (2015), for example, propose tests of the additivity restriction and develop a latent class model of non-random worker movement, while Hagedorn et al. (2017) leverage structural assumptions and long-run time series variation to estimate worker and firm effect ranks. Most recently, Bonhomme et al. (2017) show how to accommodate discrete heterogeneity and Markovian job search patterns with certain forms of endogeneity. To my knowledge no paper has yet to study mover designs in a treatment effects framework, though in applications researchers sometimes appeal to the essential logic of parallel trends.[2]

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), and Athey and Imbens (2006) – to settings with multiple non-absorbing treatment states. A related recent literature considers the treatment effect interpretations of so-called "event study" designs, in which individuals select into a binary treatment over multiple time periods (Imai and Kim, 2016; Borusyak and Jaravel, 2016; Abraham and Sun, 2018; Callaway and Sant'Anna, 2018; De Chaisemartin and D'Haultfoeuille, 2018). Here I focus on issues raised by selection into and out of multiple treatments over two periods, though an appendix section extends the theory to multiple-period mover designs. Further afield, the conditional homogeneity restriction I propose is similar to those used by Angrist and Fernandez-Val (2013), Angrist and Rokkanen (2015), and Hull (2018) for extrapolating treatment effects within and across quasi-experimental instrumental variable and regression discontinuity designs.

More generally, this paper builds on the long and rich 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 over time. Notably, the parallel trends assumptions I develop here are weaker than the "time ignorability" restrictions underlying recent approaches to non-separable panel identification (Hahn, 2001; Wooldridge, 2005; Chernozhukov et al., 2013), as the special structure of mover designs permits particular types of heterogeneous time-varying shocks.[3] As mentioned above some of my proposed estimators also allow for persistent effects of treatment – a feature that, as Imai and Kim (2016) point out, is typically ruled out in panel data frameworks.

[2] For example, Finkelstein et al. (2016)) state their identifying assumption as the restriction that "trends do not vary systematically with the migrant's origin and destination."

[3] Graham and Powell (2012) consider an alternative panel approach with continuous treatment variables, this theory is less relevant for mover selection among discrete unordered alternatives.

The remainder of this paper is organized as follows. The next section develops the dynamic treatment effects framework, characterizes the causal content of conventional mover regressions, and builds intuition for the main identification result. Section 3 then develops the class of two-step estimators for mover average treatment effects and discusses both specification testing and efficiency. Section 4 concludes. Future drafts of this paper will illustrate the theory using the the Medicare patient mover design of Finkelstein et al. (2016). All proofs, along with the extension of the theory to multiple time periods, are contained in the appendix.

2 Interpreting Mover Regressions

Suppose we observe a panel of individuals i over time periods $t = 0, \dots, T-1$, including an outcome Y_{it}, a vector of covariates X_{it}, and an individual's repeated selection $J_{it} \in \{0, \dots, J-1\}$ among J possible treatments. Let $D_{ijt} = \mathbf{1}[J_{it} = j]$ be an indicator for individual i choosing treatment j in time t. A typical mover analysis estimates the 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 individual and time fixed effects, while the coefficient β_j is meant to capture the effect of treatment j relative to the omitted treatment 0. This *mover regression* may be estimated just on the set of individuals with $J_{is} \neq J_{it}$ for some $s \neq t$ (who I refer to as *movers*) or include other *stayers* with $J_{it} = \bar{J}_i$ for all t. In practice researchers often exclude stayers or include mover-specific fixed effects in (1) in order to weaken the identifying assumptions (Finkelstein et al., 2016; Chetty and Hendren, Forthcoming).[4]

I use a dynamic potential outcomes framework (Robins, 1986, 1997) to characterize the causal interpretation of equation (1). Let $Y_{it}^{k \to j}$ denote the outcome of individual i in time t if she were to select treatment j in that period after previously following the treatment path of $k = (k_0, \dots, k_{t-1})'$. These are well-defined random variables under the usual stable unit treatment value assumption (Rubin, 1980), which I maintain throughout. When not ambiguous I write $Y_{it}^j \equiv Y_{it}^{(J_{i0}, \dots, J_{i,t-1})' \to j}$ as the time-t potential outcome of individual i given her treatment choices J_{is} for $s < t$: that is, Y_{it}^j implicitly conditions on choices made in past periods. Point-in-time treatment effects relative to the

[4] Mover regressions are often written $Y_{it} = \alpha_i + \tau_t + \beta_{J_{it}} + X_{it}'\gamma + \epsilon_{it}$. One treatment category is always omitted from estimation, though sometimes the estimated β_j are recentered to capture effects relative to the average j.

omitted treatment are written $Y_{it}^{j} - Y_{it}^{0}$, and realized outcomes can be written

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

where $\overline{0}$ denotes a conforming vector of zeros. The first two terms of equation (2) reflect potential outcomes at time t if an individual had stayed in treatment 0 in all previous periods $s < t$, while the third term captures differences in outcomes at time t arising from different treatment histories.

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

Assumption IO (*Impersistent outcomes*): For all j, t, and k, $P(Y_{it}^{k\to j} = Y_{it}^{\overline{0}\to j}) = 1$

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

Assumption CO (*Conditional orthogonality*): $E[D_{ijt}\epsilon_{it}] = 0$ for each j, where ϵ_{it} denotes the residual from the population projection of Y_{it}^{0} on X_{it} and individual and time effects

Potential outcomes are impersistent under Assumption IO in that they only depend on the contemporaneous treatment status and not on previous treatment choices. When this is the case the last term of equation (2) is ignorable. When furthermore each of the period-specific treatment effects are constant across individuals (Assumption CE), we may write $Y_{it} = Y_{it}^{0} + \sum_{j\neq 0}\overline{\beta}^{j}D_{ijt}$. The regression coefficients β_j then coincide with the causal effects $\overline{\beta}^{j}$ if we can decompose $Y_{it}^{0} = \alpha_i + \tau_t + X_{it}'\gamma + \epsilon_{it}$, with ϵ_{it} orthogonal to the vector of treatment choices (Assumption CO).[5]

Assumptions IO, CE, and CO are straightforward to state mathematically and may, as in Finkelstein et al. (2016), stem from an underlying economic model. Nevertheless, they may also prove strong and difficult to evaluate in practice: researchers might be reluctant to rule out any forms of outcome persistence or treatment effect heterogeneity, or find it challenging to assess the appropriateness of the conditional orthogonality restriction in different settings. Instead, researchers often motivate mover regressions with claims on the comparability of outcome trends across different types of movers, and validate their estimates with the kinds of pre- and post-trend analyses typically associated with difference-in-difference designs.[6] I therefore next consider what mover re-

[5] Sometimes mover regressions are motivated by a stronger conditional independence assumption, along with implicit impersistence and constant effects assumptions: see, e.g., equations (2) and (4) in Abowd et al. (2015).

[6] See, e.g., Figure 5 in Card et al. (2013), Table 4 in Jackson (2013), and Figure 6 in Finkelstein et al. (2016).

gressions identify when Assumptions IO and CE are relaxed and when Assumption CO is replaced with a quasi-experimental parallel trends assumption. To start simply and build intuition for later identification results, I first consider mover regressions with only two treatment states.

2.1 Binary Treatment Mover Regressions

Suppose there are only time periods ($T = 2$). Then the mover treatment coefficients are equivalently defined by a first-differenced regression,

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

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 a simple algebraic expression for the single mover regression coefficient β_1:

Lemma 1: If $T = J = 2$ and $X_{it} = 0$, 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 $\omega \in [0, 1]$ is a function of $P(\Delta D_{i1} = 1)$ and $P(\Delta D_{i1} = -1)$. When $P(\Delta D_{i1} = 0) = 0$, moreover, $\omega = 1/2$ and

$$\beta_1 = (E[\Delta Y_i \mid \Delta D_{i1} = 1] - E[\Delta Y_i \mid \Delta D_{i1} = -1])\omega. \tag{5}$$

The proof of Lemma 1 uses omitted-variables bias algebra to write β_1 as a linear combination of the coefficients from a saturated model for $E[\Delta Y_i \mid \Delta D_{i1}]$, identifying mean outcome growth among stayers (with $\Delta D_{i1} = 0$), those who move out of treatment 1 (with $\Delta D_{i1} = -1$), and those who move into treatment 1 (with $\Delta D_{i1} = 1$). Lemma 1 thus shows that the simplest mover regression identifies a convex average of outcome growth comparisons between movers and stayers, across the two mover types. In the special case of no stayers this expression simplifies to the single comparison (5) across the two mover groups.

Using Lemma 1, it is straightforward to show that a restriction on average outcome persistence, combined with a standard parallel trends assumption, renders binary treatment mover regressions causally interpretable:

Proposition 1: If $T = J = 2$, $X_{it} = 0$, and for each $j \in \{0, 1\}$ potential outcomes satisfy

$$E[Y_{i1}^{(1-j) \to j} \mid \Delta D_{ij} = 1] = E[Y_{i1}^{j \to j} \mid \Delta D_{ij} = 1] \tag{6}$$

and

$$E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid D_{ij0}D_{ij1} = 1] = E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid \Delta D_{ij} = 1] \tag{7}$$

$$= E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid \Delta D_{ij} = -1], \tag{8}$$

then the mover regression coefficient identifies

$$\beta_1 = \sum_{t\in\{0,1\}} \sum_{d\in\{-1,1\}} E[Y_{it}^{1} - Y_{it}^{0} \mid \Delta D_{i1} = d]\, \omega_{td}, \tag{9}$$

where $\omega_{td} \geq 0$ is a function of the distribution of $(D_{i10}, D_{i11})'$ and $\sum\limits_{t\in\{0,1\}} \sum\limits_{d\in\{-1,1\}} \omega_{td} = 1$.

In words, Proposition 1 states that the binary treatment mover regression identifies a convex combination of average treatment effects, across time and the two mover groups, under two assumptions. First, equation (6) requires individuals that move into each treatment to have, on average, the same time-1 outcome as if they had always been there (an impersistence assumption, weakening Assumption IO). Second, equations (7)-(8) state that – conditional on an individual being in treatment j at any point – the potential outcomes for different types of movers and stayers would have followed the same average growth path in the absence of a move (a parallel trends assumption, weakening Assumption CO). Note that Proposition 1 does not directly restrict treatment effect heterogeneity, relaxing Assumption CE.

Intuition for the link between equation (4) and equation (9) comes from classic difference-in-differences logic. Under parallel trends, the difference in outcome growth rates between those moving into treatment 1 at $t = 1$ (with $\Delta D_{i1} = 1$) and treatment 0 stayers (with $D_{i00} = D_{i01} = 1$) identifies the average time-1 treatment effect of the former group,

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

while the average time-1 treatment effect for the other mover group (with $\Delta D_{i1} = -1$) is identified by subtracting their outcome growth from that of the other stayer group (with $D_{i10} = D_{i11} = 1$):

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

Similarly, when the assumptions of Proposition 1 hold, comparisons of outcome growth among (i) $\Delta D_{i1} = 1$ movers and treatment 1 stayers and (ii) treatment 0 stayers and $\Delta D_{i1} = -1$ movers identify average time-0 treatment effects. Thus each of the two terms in Lemma 1 can be written as a weighted average of difference-in-difference comparisons identifying average causal effects under the assumptions.

Figure 1(a) summarizes the four difference-in-difference comparisons combined in the simple mover regression. Outcome growth contrasts within the dark- and light-colored groups identify average time-1 effects under the assumptions of Proposition 1, while comparisons within the dashed- and solid-lines group identify average time-0 effects. Interestingly, only the parallel trends assumption is needed to identify the former, whereas the impersistence restriction is also used to ensure the time-0 comparisons are causally interpretable. This is because time-0 outcomes can always be "differenced off" when comparing within the light- dark-colored groups of Figure 1(a), as in standard difference-in-differences, but the time-1 outcomes in the "reverse" difference-in-differences of the dashed- and solid groups are not comparable when potential outcomes systematically persist.

The appendix proof to Proposition 1 shows that the weights aggregating these difference-in-difference comparisons depend on the population shares of different mover and stayer types. Clearly if there are no movers with $\Delta D_{i1} = -1$ then β_1 represents a weighted average of causal effects for movers 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 each treatment: at the extremes if there are no stayers at 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 when there are no initially-treated individuals, the mover regression identifies $E[Y_{i1}^1 - Y_{i1}^0 \mid \Delta D_{i1} = 1] = E[Y_{i1}^1 - Y_{i1}^0 \mid D_{i11} = 1]$, the average treatment effect on the treated, as in a standard difference-in-difference design.

Finally, it is worth exploring the special situation in which there are no stayers of either type. This follows from the second part of Lemma 1:

Corollary to Proposition 1: Suppose $P(\Delta D_{i1} = 0) = 0$ and the assumptions of Proposition 1 hold. Then the mover regression coefficient identifies

$$\beta_1 = \frac{1}{2}(E[Y_{i1}^1 - Y_{i1}^0 \mid \Delta D_{i1} = 1] + E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} = -1]) \tag{12}$$

$$= \frac{1}{2}(E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} = 1] + E[Y_{i1}^1 - Y_{i1}^0 \mid \Delta D_{i1} = -1]). \tag{13}$$

As mentioned, researchers may exclude stayers from a mover regression in order to weaken the identifying assumptions: here note that whenever the parallel trends assumption in Proposition is satisfied with stayers it is also satisfied when they are excluded. The corollary shows that without stayers the simple mover regression identifies a simply-weighted average of mover treatment effects across time. Interestingly, this estimand can be expressed in two ways: as the average of time-t effects for movers into treatment 1 and time-$(1-t)$ effects for movers out of treatment 1, for either $t = 0$ or $t = 1$. The equivalence of (12) and (13) is an algebraic consequence of imposing parallel trends for both treatments.

A pessimistic interpretation of Proposition 1 and its corollary is that mover regression coefficients may be difficult to interpret even in the binary treatment case. That is, two mover experiments with the same joint distribution of causal effects and treatment choices (and thus the same average treatment effects for different types of movers and stayers) may produce different regression coefficients, depending on the marginal distribution of initial treatment. Researchers interested in the external validity of mover regression estimates, or in comparing estimates across different experiments, may view this as an important limitation.[7] Nevertheless, the above discussion suggests this limitation can be easily overcome: when the assumptions of Proposition 1 hold, a researcher wishing to estimate a time-specific average causal effect could construct the relevant difference-in-difference comparisons in Figure 1(a) and weight them together as they please. This reweighted difference-in-difference logic is at the core of the general strategy for identifying mover average treatment effects in Section 3. Before formalizing the strategy, however, we first consider additional complications arising from mover designs with multiple unordered treatments.

2.2 Multiple-treatment Mover Regressions

In practice, mover regressions involve choices across many different treatments. Finkelstein et al. (2016), for example, estimate models with 305 healthcare market treatments, while Card et al. (2013) study worker movement across over a million German firms. It may be reasonable to expect that the basic difference-in-difference logic extends to multiple-treatment regressions, so that they again capture some weighted average of causal effects under weak quasi-experimental restrictions. The following result, however, shows that this is not the case:

Proposition 2: Suppose $T = 2$, $X_{it} = 0$, and $J > 2$. Then even if both Assumption IO and the parallel trends assumption in Proposition 1 hold for all treatments j, the mover regression coefficients need not identify weighted averages of individual treatment effects.

The appendix proof of Proposition 2 shows that even with strongly impersistent outcomes and conventionally parallel trends, the multiple treatment coefficients in equation (1) combine a set of average treatment effects with a set of non-causal terms. The latter are of the form

$$E[Y_{i1}^j - Y_{i0}^j \mid D_{ij0}D_{ij1} = 1] - E[Y_{i1}^k - Y_{i0}^k \mid D_{ik0}D_{ik1} = 1] \tag{14}$$

and

$$E[Y_{it}^j - Y_{it}^0 \mid D_{i00}D_{ij1} = 1] + E[Y_{is}^k - Y_{is}^j \mid D_{ij0}D_{ik1} = 1] - E[Y_{iu}^k - Y_{iu}^0 \mid D_{i00}D_{ik1} = 1], \tag{15}$$

[7] Yitzhaki (1996) argues this point for regression-weighted averages of causal parameters in general; see also Angrist (1998) and, more recently, Gibbons et al. (2018).

for treatments $j > 0$ and $k \neq j$ and for time periods t, s, and u. These capture, respectively, differences in trends among stayers in different treatment states and combinations of causal effects for movers across different treatments and times. Although each of these comparisons are ignorable in the canonical additively-separable and constant-effects model (or imposed directly by Assumption CO and CE), in general they are non-zero under the parallel trends assumption, rendering the mover coefficients causally uninterpretable.

At first this shortcoming of multi-treatment mover regressions may appear a puzzle. After all additional treatment choices simply yield more difference-in-difference comparisons, each of which are causally interpretable under parallel trends and Assumption IO. The issue, as with the weighting scheme of Proposition 1, is that the mover regression combines the set of quasi-experiments in a way that is sensible when the canonical mover model is correctly specified, but need not be when there are heterogeneous treatment effects or time-varying shocks.

As a simple example of the issue, consider a three-treatment design with the two mover and stayer groups illustrated in Figure 1(b). Following the logic of the previous subsection, there are two time-1 average causal effects identified under parallel trends: the effect of treatment 1 relative to treatment 0 for movers from 0 to 1 and the effect of treatment 2 relative to treatment 1 for movers from 1 to 2. The average time-0 effect of treatment 1 relative to treatment 0 is moreover identified for movers from 0 to 1 if we add an outcome impersistence condition. As in Figure 1(a), the difference-in-difference comparisons identifying each of these effects are given by contrasts within similarly-colored and similarly-patterned line groups.

How does the mover regression combine these effects? Using formulas from the proof to Proposition 2, we can show that the treatment coefficients satisfy

$$\begin{aligned}
\beta_1 =& E[Y_{i1}^1 - Y_{i1}^0 \mid \Delta D_{i1} = 1]p_0 + E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} = 1](1 - p_0) && (16)\\
\beta_2 =& E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} = 1] + E[Y_{i1}^2 - Y_{i1}^1 \mid \Delta D_{i2} = 1] && (17)\\
&+ 2p_0 \left(E[Y_{i1}^1 - Y_{i0}^1 \mid D_{i10}D_{i11} = 1] - E[Y_{i1}^0 - Y_{i0}^0 \mid D_{i00}D_{i01} = 1] \right),
\end{aligned}$$

where p_0 denotes the proportion of stayers in treatment 0. Thus, while the treatment-1 mover regression coefficient continues to identify a convex average of time-0 and time-1 treatment effects, the coefficient on treatment 2 is not causal. Lacking any difference-in-difference comparison identifying the effect of treatment 2 relative to treatment 0, the mover regression sums the first two causal effects in equation (17). Under a constant effects assumption this is sensible, since then

$$\begin{aligned}
E[Y_{i0}^1 - Y_{i0}^0 \mid \Delta D_{i1} = 1] + E[Y_{i1}^2 - Y_{i1}^1 \mid \Delta D_{i2} = 1] &= \overline{\beta}^1 + (\overline{\beta}^2 - \overline{\beta}^1) = \overline{\beta}^2 \\
&= E[Y_{i1}^2 - Y_{i1}^0 \mid \Delta D_{i2} = 1], \qquad (18)
\end{aligned}$$

though with meaningful treatment effect heterogeneity this sum need not be causally interpretable. Equation (17) moreover includes a term capturing the difference in outcome growth rates for the two stayer groups. This is again sensible under Assumptions CE and CO, since then

$$\begin{aligned} E[Y^1_{i1} - Y^1_{i0} \mid D_{i10}D_{i11} = 1] - E[Y^0_{i1} - Y^0_{i0} \mid D_{i00}D_{i01} = 1] =& E[\overline{\beta}^1 + Y^0_{i1} - \overline{\beta}^1 + Y^0_{i0} \mid D_{i10}D_{i11} = 1] \\ &- E[Y^0_{i1} - Y^0_{i0} \mid D_{i00}D_{i01} = 1] \\ &= 0, \end{aligned} \tag{19}$$

though in general this term need not be zero. Note that the left-hand side of equation (18) is an example of equation (15), where the third undefined term is arbitrarily set to zero, while the left-hand side of equation (19) is an example of equation (14).

A researcher pessimistic about the weighting scheme in Proposition 1 may therefore have even more cause for pessimism with multiple-treatment mover regressions. Despite the availability of multiple difference-in-difference comparisons, conventional mover analyses need not even have a weighted causal effect interpretation when $J > 2$. Nevertheless, as with the binary treatment case, one could imagine individually extracting and more sensibly combining the component difference-in-difference quasi-experiments to overcome the limitations of mover regressions. This combination may involve extrapolating some causal effects from others, just as the above regression example does in order to identify the average effect of treatment 2 relative to treatment 0 under constant effects. Weaker extrapolations, however, may only combine difference-in-difference experiments capturing treatment effects from the same time period and for observably-similar movers, weakening the constant effects assumption. I next develop a class of two-step estimators that enact this logic.

3 Estimating Mover Average Treatment Effects

In general there may be many combinations of heterogeneous treatment effects that are of interest in a mover design. To discipline the initial theoretical approach I focus on mover average treatment effects, defined for each treatment $j > 0$ as

$$MATE_{jt} = E[Y^j_{it} - Y^0_{it} \mid \Delta D_i \neq 0], \tag{20}$$

where ΔD_i is a vector collecting the set of ΔD_{ij}. Here $MATE_{jt}$ captures the average time-t effect of treatment j relative to treatment 0, among individuals that change treatment status.[8] That mover

[8] Recall that while for simplicity we restrict attention here to two periods, the appendix generalizes what follows to the multiple period case. For this I define ΔD_i as a vector collecting the set of $D_{ijt} - D_{ijs}$ for all $t \neq s$, so that $MATE_{jt}$ captures the average effect for individuals moving at any point.

designs tend to reveal effects on mobile individuals is often implicit in applications, just as is that standard difference-in-difference estimation tends to capture average effects for those who become treated (Abadie, 2005). Of course, if there are no stayers in the study population the MATEs become average treatment effects (ATEs).

3.1 Identifying Assumptions

The key quasi-experimental assumption I leverage is that of conditional parallel trends, which generalizes the trend restrictions in Proposition 1. Formally, with X_i denoting a vector of controls (including, perhaps, some elements of the X_{i0} and X_{i1} from the mover regression and other time-invariant observables), consider

Assumption CPT (*Conditional parallel trends*): For each treatment j and x in the support of X_i,

$$\begin{aligned} E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid D_{ij0}D_{ij1} = 1, X_i = x] &= E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid \Delta D_{ij} = 1, X_i = x] && (21) \\ &= E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid \Delta D_{ij} = -1, X_i = x]. && (22) \end{aligned}$$

Under Assumption CPT, the average treatment-j outcomes for different types of movers into or out of j and stayers at j would have followed parallel paths if not for the move, conditional on the controls in X_i. 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 pre-treatment income for those entering job training programs (Ashenfelter, 1978; Ashenfelter and Card, 1985). Clearly, the parallel trends assumption in Proposition 1 is a special case of Assumption CPT, for which $X_i = 0$. It is also straightforward to verify that the "time ignorability" identifying assumptions developed in the recent literature on non-separable panel models (e.g. Chernozhukov et al. (2013)) imply Assumption CPT, but are not implied by it.[9]

As with multiple-treatment mover regressions, combining causal effects across many difference-in-difference experiments requires further homogeneity restrictions. Here I adopt a weaker assumption than the canonical model: that mover treatment effects are on average comparable, conditional on observables:

Assumption CEH (*Conditional effect homogeneity*): For each period t, treatments j and k, and x in the support of X_i, $E[Y_{it}^{j} - Y_{it}^{k} \mid \Delta D_i = d, X_i = x]$ does not depend on $d \neq 0$.

[9] Chernozhukov et al. (2013) consider models of the form $Y_{it}^{k\to j} = g(j, X_i, \alpha_i, \epsilon_{it})$, where the distribution of ϵ_{it} does not depend on t given $(\alpha_i, J_{i0}, J_{i1}, X_i)$. Then $E[Y_{i1}^{j\to j} - Y_{i0}^{j} \mid J_{i0}, J_{i1}, X_i] = 0$, satisfying Assumption CPT. Identification here allows for heterogeneous time-varying shocks by leveraging the particular structure of mover designs.

Restrictions on the conditional heterogeneity of average causal effects have been previously used in the treatment effects literature to extrapolate within and across different quasi-experiments (Angrist and Fernandez-Val, 2013; Angrist and Rokkanen, 2015; Hull, 2018). Here Assumption CEH states that differences in average causal effects for movers with different origins-destination pairs are driven only by the set of observed contemporaneous or lagged controls in X_i. In applications researchers sometimes gauge mover effect homogeneity by tests of outcome trend symmetry (e.g. Card et al. (2013)); Assumption CEH can thus be thought to relax the assumptions motivating such tests.

Finally, as in the previous section, I use a restriction on average potential outcome persistence to identify time-0 causal effects from "reverse" difference-in-difference quasi experiments. Again this can be made conditional on observables:

Assumption COI (*Conditional outcome impersistence*): For all treatments (j, k) and $x \in Supp(X_i)$,

$$E[Y_{i1}^{k\to j} \mid D_{ik0}D_{ij1} = 1, X_i = x] = E[Y_{i1}^{j\to j} \mid D_{ik0}D_{ij1} = 1, X_i = x]. \tag{23}$$

Under Assumption COI, the mean outcome for movers into each treatment j from each treatment k is the same as it would be if the movers had always chosen j, conditional on the controls in X_i. Thus any potential for persistence in average outcomes for movers must be driven by time-varying or invariant observables, relaxing the usual panel data restriction of complete outcome impersistence (Imai and Kim, 2016).

Together, these three assumptions relax those of standard mover regressions. In particular Assumptions CPT and CEH are implied by a partially-separable model of dynamic potential outcomes,

$$Y_{i0}^{j} = \alpha_i + \beta_{j0}(X_i) + \epsilon_{i0} \tag{24}$$

$$Y_{i1}^{k\to j} = \alpha_i + \beta_{j1}(X_i) + \epsilon_{ik1}, \tag{25}$$

where $\epsilon_{ij1} - \epsilon_{i0}$ is mean-independent of treatment choices conditional on the controls and on i being a j-mover or j-stayer (that is, of (J_{i0}, J_{i1}) conditional on X_i and $D_{ij0} + D_{ij1} \neq 0$). As shown below, with enough stayers this model permits identification of time-1 MATEs, here taking the form

$$MATE_{jt} = E[\beta_{jt}(X_i) - \beta_{0t}(X_i) \mid \Delta D_i \neq 0]. \tag{26}$$

Estimating time-0 MATEs or omitting stayers will additionally require Assumption COI, which here restricts $E[\epsilon_{ik1} \mid J_{i0}, J_{i1}, X_i] = E[\epsilon_{im1} \mid J_{i0}, J_{i1}, X_i]$ for all k, m. In contrast, the mover regression assumptions IO, CE, and CO are satisfied when the ϵ_{ikt} do not depend on k and are mean-independent of (J_{i0}, J_{i1}), and when $\beta_{jt}(X_i)$ is additively-separable in treatment, time, and a fixed linear combination of the controls. The following identification results thus allow for both time-varying shocks and treatment effect heterogeneity to vary flexibly with observables.

3.2 Identification with and without Stayers

With the three key assumptions in hand, I next establish MATE identification in two salient cases. First, I suppose a researcher is willing to assume potential outcome trends are conditionally comparable between movers and stayers, and that stayers are sufficiently dispersed to make feasible a set of conditional difference-in-difference comparisons linking treatments j and 0. Enumerating this set requires some additional notation; I let C_j denote the set of all variable-length n-tuples $C_{j\ell} = (c_{j\ell 0}, \dots, c_{j\ell M_{j\ell}})$, where $M_{j\ell} + 1$ denotes the length of $C_{j\ell}$ and where $c_{j\ell m} \in \{0, \dots, J-1\}$, $c_{j\ell 0} = 0$, $c_{j\ell M_{j\ell}} = j$, and $c_{j\ell m} \neq c_{j\ell n}$ for all $m \neq n$. Thus C_j collects all of the ℓ possible paths (or *chains*) from treatment 0 to treatment j via other treatment states, where no intermediate treatment is included as a link in the chain more than once. We then have the following result:

Proposition 3: Suppose $T = 2$, Assumptions CPT and CEH hold, and that for treatment j and period t there exists a chain $C_{j\ell}$ such that, for each $m = 1, \dots, M_{j\ell}$ and all x in the support of X_i either (i) $P(D_{i,m-1,1-t}D_{imt} = 1 \mid X_i = x) > 0$ and $P(D_{i,m-1,0}D_{i,m-1,1} = 1 \mid X_i = x) > 0$ or (ii) $P(D_{im,1-t}D_{i,m-1,t} = 1 \mid X_i = x) > 0$ and $P(D_{im0}D_{im1} = 1 \mid X_i = x) > 0$. Then, for $t = 1$,

$$MATE_{jt} = E\left[\Delta Y_i \left(\sum_{m=1}^{M_{j\ell}} (w_{j\ell m}\rho_{it}^{c_{j\ell,m-1},c_{j\ell m}} + (1 - w_{j\ell m})(-\rho_{it}^{c_{j\ell m},c_{j\ell,m-1}}))\right)\right], \tag{27}$$

where the $w_{j\ell m}$ are constants such that $w_{j\ell m} = 0$ if (i) fails for m and $w_{j\ell m} = 1$ if (ii) fails for m, and where

$$\rho_{it}^{c,d} = (-1)^t D_{ic,1-t} \frac{D_{ict}E[D_{idt}D_{ic,1-t} \mid X_i] - D_{idt}E[D_{ict}D_{ic,1-t} \mid X_i]}{E[D_{ict}D_{ic,1-t} \mid X_i]E[D_{idt}D_{ic,1-t} \mid X_i]} \frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)}. \tag{28}$$

If moreover Assumption COI holds, equations (27) and (28) also hold for $t = 0$.

In words, Proposition 3 states that the time-1 mover average treatment effect for treatment j is identified by a particular weighted average of outcome growth ΔY_i under conditional parallel trends and effect homogeneity, and when there exists a set of difference-in-difference comparisons linking treatment j to the reference treatment 0. This set is given by the chain $C_{j\ell} = (0, c_{j\ell 1}, \dots, c_{j\ell, M_{j\ell}-1}, j)$, where between any two links $c_{j\ell,m-1}$ and $c_{j\ell m}$ there exist either (i) movers from treatment $c_{j\ell,m-1}$ into treatment $c_{j\ell m}$ and stayers at treatment $c_{j\ell,m-1}$ or (ii) movers from treatment $c_{j\ell m}$ into treatment $c_{j\ell,m-1}$ and stayers at treatment $c_{j\ell m}$, conditional on the controls. The proof to Proposition 3 shows that under these assumptions the m-specific terms in the weighting scheme (27) identify $E[Y_{i1}^{c_{j\ell,m}} - Y_{i1}^{c_{j\ell,m-1}} \mid \Delta D_i \neq 0]$, so that summing over the links of the chain identifies $E[Y_{i1}^j - Y_{i1}^0 \mid \Delta D_i \neq 0] = MATE_{j1}$. An analogous result follows for $t = 0$ effects when potential outcomes are conditionally impersistent.

Unpacking this result further, note that each summand of the weighting scheme (27) is in turn a linear combination of two terms given by equation (28). The $\rho_{it}^{c,d}$ depend on the conditional frequency of different groups of movers and stayers given the controls, as summarized by the propensity scores $E[D_{idt}D_{ics} \mid X_i]$ and $P(\Delta D_i \mid X_i)$. The appendix proof shows that weighting ΔY_i by $\rho_{it}^{c,d}$ and $-\rho_{it}^{d,c}$ replicates and averages together conditional difference-in-difference comparisons between the different mover and stayer groups illustrated in Figure 1(a). For example,

$$E[\Delta Y_i \rho_{i1}^{c,d} \mid X_i] = (E[\Delta Y_i \mid D_{ic0}D_{id1} = 1, X_i] - E[\Delta Y_i \mid D_{ic0}D_{ic1} = 1, X_i])\frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)}, \quad (29)$$

which is a weighted comparison of the conditional difference in outcome growth between movers from treatments c to d and stayers at treatment c. Similarly, $E[\Delta Y_i(-\rho_{i1}^{d,c}) \mid X_i]$ identifies a weighted conditional contrast of outcome growth between stayers at treatment d and movers from treatments d to c. As before when the conditional parallel trends assumption holds these identify conditional weighted average treatment effects of the respective mover groups, which are assumed to both be representative of all movers under conditional effect homogeneity. Thus any weighted average of $E[\Delta Y_i \rho_{i1}^{c,d} \mid X_i]$ and $E[\Delta Y_i(-\rho_{i1}^{d,c}) \mid X_i]$ identifies $E[Y_{i1}^{c_{j\ell,m}} - Y_{i1}^{c_{j\ell,m-1}} \mid \Delta D_i \neq 0, X_i]\frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)}$ under the assumptions; averaging these averages over the marginal distribution of the controls then identifies $E[Y_{i1}^{c_{j\ell,m}} - Y_{i1}^{c_{j\ell,m-1}} \mid \Delta D_i \neq 0]$.

Proposition 3 can be thought to generalize Abadie (2005)'s approach to identification in difference-in-difference designs. Specifically, suppose $J = 2$ and there are no movers from treatment 1 to treatment 0. Then there is only one chain $C = (0, 1)$ satisfying (i), with $w_1 = 1$ and Assumption CEH satisfied trivially, and the weighting scheme identifying $MATE_{11} = E[Y_{i1}^1 - Y_{i0}^0 \mid D_{i1} = 1]$ coincides with that of Abadie (2005).

It is also worth noting that the logic of Proposition 3 implies a weaker approach to MATE identification in binary treatment mover designs, using particular data-driven weights to avoid restricting treatment effect heterogeneity:

Corollary to Proposition 3: Suppose $J = T = 2$, Assumption CPT holds, and for period t and the chain $C = (0, 1)$ either condition (i) or (ii) from Proposition 3 holds. Then without imposing Assumption CEH we have, for $t = 1$,

$$MATE_{1t} = E\left[\Delta Y_i \left(w_{it}^* \rho_{it}^{0,1} + (1 - w_{it}^*)(-\rho_{it}^{1,0})\right)\right], \quad (30)$$

where

$$w_{it}^* = \frac{P(D_{i1t}D_{i0,1-t} = 1 \mid X_i)}{P(\Delta D_{i1} \neq 0 \mid X_i)} \quad (31)$$

If moreover Assumption COI holds, equations (30) and (31) also hold for $t = 0$.

As in Proposition 3, the weighting scheme in equation (30) combines two sets of conditional difference-in-difference experiments via the $\rho_{it}^{0,1}$ and $(-\rho_{it}^{1,0})$ terms. Here however these terms are combined via weights (31) that are proportional to the share of movers of each type – either out of or into treatment 1 – in order to avoid mixing effects across the two groups. This corollary thus gives a flexible way to estimate pairwise average treatment effects in a mover design for movers across two treatments, though such effects need not be comparable across different treatments without an effect homogeneity assumption.

To apply Proposition 3 or its corollary, a researcher must be willing to assume that any systematic differences in the potential outcome trends of movers and stayers arise from a set of observable controls. As in the previous section, we may also consider identification under a weaker parallel trends assumption in which the trends of stayers are unrestricted. The following result shows that in such cases one may still identify the average of $MATE_{jt}$ across time t:

Proposition 4: Suppose $T = 2$, Assumptions CPT, CEH, and COI hold, and for treatment j there exists a chain $C_{j\ell}$ such that, for each $m = 1, \ldots, M_{j\ell}$ and all x in the support of X_i, $P(D_{i,m-1,0}D_{im1} = 1 \mid X_i = x) > 0$ and $P(D_{im0}D_{i,m-1,1} = 1 \mid X_i = x) > 0$. Then

$$\frac{1}{2}(MATE_{j0} + MATE_{j1}) = E\left[\Delta Y_i \left(\sum_{m=1}^{M_{j\ell}} \kappa_i^{c_{j\ell,m-1},c_{j\ell m}}\right)\right], \tag{32}$$

where

$$\kappa_i^{c,d} = \frac{1}{2}\frac{D_{ic0}D_{id1}E[D_{id0}D_{ic1} \mid X_i] - D_{id0}D_{ic1}E[D_{ic0}D_{id1} \mid X_i]}{E[D_{id0}D_{ic1} \mid X_i]E[D_{ic0}D_{id1} \mid X_i]}\frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)}. \tag{33}$$

Proposition 4 generalizes the basic logic of the corollary to Proposition 1, along the lines of Proposition 3. Here each summand in equation (32) identifies a weighted difference-in-difference comparison between movers from treatment c to treatment d and movers from treatment d to treatment c, such that, under the assumptions,

$$\begin{aligned} E[\Delta Y_i \kappa_i^{c,d} \mid X_i] &= \frac{1}{2}(E[\Delta Y_i \mid D_{ic0}D_{id1} = 1, X_i] - E[\Delta Y_i \mid D_{id0}D_{ic1} = 1, X_i])\frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)} \\ &= \frac{1}{2}(E[Y_{i0}^d - Y_{i0}^c \mid \Delta D_i \neq 0, X_i] + E[Y_{i1}^d - Y_{i1}^c \mid \Delta D_i \neq 0, X_i])\frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)}. \end{aligned} \tag{34}$$

Similar to before, weighting and adding together these comparisons along the links of any chain C_j from treatment 0 to treatment j thus identifies $\frac{1}{2}(MATE_{j0} + MATE_{j1})$. For this to be feasible there must exist both types of movers at each link, conditional on the controls, as no stayer variation is used. The cost of the weaker parallel trends assumption is a somewhat less informative estimand: the weighting scheme (32) is not able to separately identify time-0 and time-1 effects.

3.3 Estimation and Testing

As in Abadie (2005), Propositions 3 and 4 suggest a straightforward two-step approach to estimating mover average treatment effects from an *i.i.d.* sample of size N. In a first step, a researcher computes a set of propensity score estimates $\widehat{E}[D_{id0}D_{ic1} \mid X_i]$, along with the sample proportion of movers $\widehat{P}(\Delta D_i \neq 0)$. For a given chain $C_{j\ell}$, she then forms the sample analogue of either equation (27) or equation (32), with the former also requiring specification of a vector of weights $w_{j\ell}$. When the first-step estimates are consistent, so too will be the second-step weighting estimator under the assumptions of Proposition 3 or 4. The asymptotic behavior of this estimator depends on the properties of the data-generating process and whether or not the first-step propensity score estimates are parametric. Future drafts will establish this behavior formally using standard first-order asymptotic theory (as in Abadie (2005)); for now I take the $\sqrt{N}$-consistency and asymptotic normality of the described two-step estimators as given, in order to focus on additional conceptual issues arising from overidentification in mover designs.

When both conditions (i) and (ii) of Proposition 3 hold for some of the links of a chain $C_{j\ell}$, one faces the choice of which set of weights $w_{j\ell m}$ to use for estimating $MATE_{jt}$. Further degrees of overidentification arise when multiple $C_{j\ell}$ satisfy the assumptions of either Propositions 3 or 4: in general the number of possible chains grows rapidly with the number of institutions, though in practice the graph connecting any two institutions by difference-in-difference experiments may be sparse.[10] Testing the equality of two estimates formed by different chains or weights constitutes an omnibus specification test of the identifying assumptions which may be used to gauge the plausibility of the quasi-experimental framework.

A potentially more powerful test of this joint null hypothesis uses a conditionally-efficient estimator, optimally combining different quasi-experiments to minimize the asymptotic estimator variance. To characterize this procedure, let $\widehat{\rho}_{it}^{c,d}$ and $\widehat{\kappa}_{it}^{c,d}$ be consistent estimates of $\rho_{it}^{c,d}$ and $\kappa_{it}^{c,d}$, and let $\widehat{M}$ denote a $K \times 1$ vector collecting the set of either all positive and negative $\widehat{E}[\Delta Y_i \widehat{\rho}_{it}^{c,d}]$ or of all $\widehat{E}[\Delta Y_i \widehat{\kappa}_{it}^{c,d}]$ that are well-defined for treatments (c, d), where $\widehat{E}[V_{it}]$ denotes the sample average of a variable V_{it}. Next, let S_j be a $P \times K$ matrix with elements $S_{jpk} \in \{0, 1\}$, such that, for all p, $\sum_k S_{jpk}\widehat{M}_k$ is consistent for $MATE_{jt}$ under Proposition 3 or identifies $\frac{1}{2}(MATE_{j0} + MATE_{j1})$ under Proposition 4. Then a general method of moments estimator for the target causal parameter

[10] To be precise, there are $\sum_{k=0}^{J-2} k!\binom{J-2}{k}^2$ possible chains given J institutions.

β_j, which combines mover variation across all available chains, is given by

$$\widehat{\beta}_j = \arg\min_{\beta_j}(\beta_j - S_j\widehat{M})'W(\beta_j - S_j\widehat{M}), \tag{35}$$

where W is some $P \times P$ positive-definite weighting matrix. Given the first-step propensity scores underlying $\widehat{M}$, a conditionally-efficient estimator $\widehat{\beta}^*_j$ is obtained by setting $W = (S_j\widehat{\Omega}S_j')^{-1}$, where $\widehat{\Omega}$ consistently estimates the asymptotic variance of $\widehat{M}$. Solving (35), we then have

$$\widehat{\beta}^*_j = \frac{\iota'(S_j\widehat{\Omega}S_j')^{-1}S_j}{\iota'(S_j\widehat{\Omega}S_j')^{-1}\iota}\widehat{M}, \tag{36}$$

where ι denotes a $P \times 1$ vector of ones.

This estimate of the target causal parameter (either $MATE_{jt}$ if $\widehat{M}$ collects moments involving $\widehat{\rho}^{c,d}_{it}$ or $\frac{1}{2}(MATE_{j0} + MATE_{j1})$ if $\widehat{M}$ collects moments involving $\widehat{\kappa}^{c,d}_{it}$) is an optimally-weighted combination of all relevant difference-in-difference quasi-experiments, with weights proportional to the elements of $\iota'(S_j\widehat{\Omega}S_j')^{-1}S_j$. In particular, it satisfies

$$\sqrt{N}(\hat{\beta}^*_j - \beta) \Rightarrow \mathcal{N}(0, (\iota'(S_j\Omega S_j')^{-1}\iota)^{-1}), \tag{37}$$

where $\widehat{\Omega} \xrightarrow{p} \Omega$. As usual, an omnibus specification test based on (36) is given by

$$\widehat{T}_j = (\beta^*_j - S_j\widehat{M})'(S_j\widehat{\Omega}S_j')^{-1}(\beta^*_j - S_j\widehat{M}), \tag{38}$$

which has an asymptotic χ^2_{P-1} distribution under the joint null of the identifying assumptions. Implicitly, (38) checks whether the same estimate of β_j is obtained from any two difference-in-difference chains, weighting pairwise comparisons by the efficient weights.

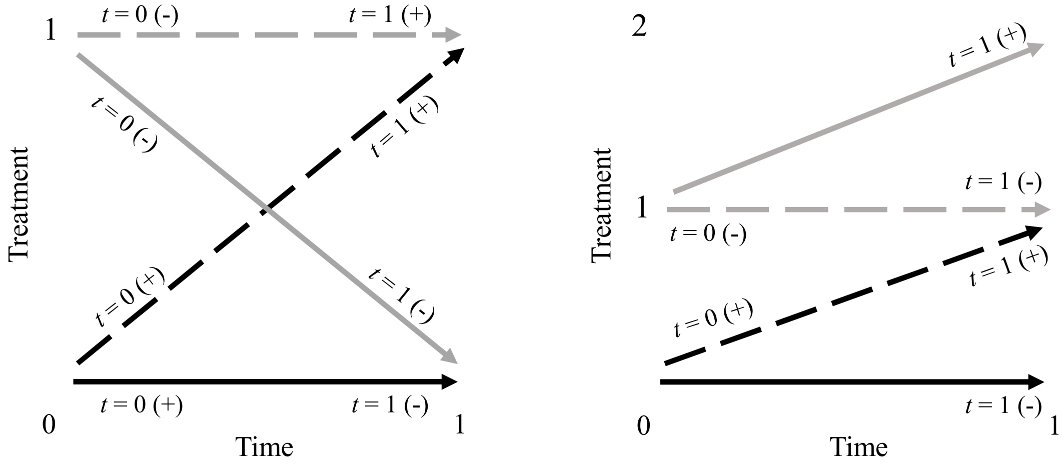
It appears relatively straightforward to compute equations (36) and (38) in practice. Estimation of the large set of mover propensity scores can be distributed over multiple computational resources, while the second-step calculation of the estimated asymptotic variance of $\widehat{M}$ and the second-step estimator (36) is likely to be quite simple. Future drafts of this paper will include simulations of the computational demands and finite-sample performance of these estimators, as well as an application to a real-world movers design.

4 Conclusions

Although retaining the flavor of simpler difference-in-differences designs, quasi-experimental mover designs require additional restrictions. Mover regression estimates, while causally interpretable in the binary treatment case, fail in general to recover weighted averages of heterogeneous treatment effects under a parallel trends assumption alone. In contrast, the two-step estimators developed here accommodate heterogeneous treatment effects and time-varying shocks, provided they are not correlated with individual movement conditional on controls. Certain mover average treatment effects are moreover identified without a direct restriction on potential outcome persistence, provided the potential trends of movers and stayers are conditionally comparable. As argued above, the computation of conditionally-efficient treatment effect estimates is likely to be 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).

As always, whether the assumptions considered here are more plausible than other approaches to mover designs , such as Hagedorn et al. (2017) and Bonhomme et al. (2017), will be a matter of context. The restrictions of conditional parallel trends and conditional effect homogeneity are likely to have the advantage of both being familiar to applied researchers and relatively straightforward to consider in applications. At a minimum, the quasi-experimental approach may allow researchers to verify the robustness of substantive conclusions drawn from conventional mover regressions and more recent variations.

Figures and Tables



(a) Binary treatment

(b) Multiple treatments

Figure 1: Example difference-in-difference comparisons in mover designs

Notes: This figure illustrates groups of movers and stayers identifying time-t causal effects when outcomes are impersistent and parallel trends holds. In each panel, time-0 effects are identified by outcome growth contrasts within the dashed- and (in Figure 1(a)) solid-line groups, while time-1 effects are identified by outcome growth contrasts within the light- and dark-colored groups. The time-specific notes on each line indicate whether the outcome growth for that subgroup is to be added or subtracted. Panel (a) shows the full set of difference-in-difference comparisons with binary treatments, while panel (b) shows the comparisons for the multiple treatment example discussed in the text.

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.

Abraham, S. and L. Sun (2018): "Estimating Dynamic Treatment Effects in Event Studies," 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.

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," Working Paper.

Borusyak, K. and X. Jaravel (2016): "Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume," Working Paper.

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.

Callaway, B. and P. H. C. Sant'Anna (2018): "Difference-in-Differences with Multiple Time Periods and an Application on the Minimum Wage and Employment," Working Paper.

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 (Forthcoming): "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects," *Quarterly Journal of Economics*.

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

de Chaisemartin, C. and X. D'Haultfoeuille (2018): "Two-way 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.

Gibbons, C. E., J. C. S. Serrato, and M. B. Urbancic (2018): "Broken or Fixed Effects?" *Journal of Econometric Methods*, 20170002.

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. (2018): "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.

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.

Molitor, D. (2017): "The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration," *American Economic Journal: Economic Policy*, 10, 326–356.

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.

Rubin, D. B. (1980): "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment," *Journal of the American Statistical Association*, 75, 591–593.

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

$$\begin{aligned} \Delta Y_i &= \tau + \beta_1 \Delta D_{i1} + \Delta\epsilon_i \\ &= \tau + \beta_1(\mathbf{1}[\Delta D_{i1} = 1] - \mathbf{1}[\Delta D_{i1} = -1]) + \Delta\epsilon_i, \end{aligned} \tag{39}$$

which is nested by the regression

$$\begin{aligned} \Delta Y_i &= \widetilde{\tau} + \widetilde{\beta}_1(\mathbf{1}[\Delta D_{i1} = 1] - \mathbf{1}[\Delta D_{i1} = -1]) + \widetilde{\alpha}_1\mathbf{1}[\Delta D_{i1} = -1] + \Delta\widetilde{\epsilon}_i \\ &= \widetilde{\tau} + \widetilde{\beta}_1\mathbf{1}[\Delta D_{i1} = 1] + (\widetilde{\alpha}_1 - \widetilde{\beta}_1)\mathbf{1}[\Delta D_{it} = -1] + \Delta\widetilde{\epsilon}_i. \end{aligned} \tag{40}$$

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

$$\widetilde{\beta}_1 = E[\Delta Y_i \mid \Delta D_{i1} = 1] - E[\Delta Y_i \mid \Delta D_{i1} = 0] \tag{41}$$

$$\text{and} \quad \widetilde{\alpha}_1 - \widetilde{\beta}_1 = E[\Delta Y_i \mid \Delta D_{i1} = -1] - E[\Delta Y_i \mid \Delta D_{i1} = 0]. \tag{42}$$

Thus, by usual omitted-variables bias logic,

$$\beta_1 = \widetilde{\beta}_1 + \widetilde{\alpha}_1 \frac{Cov(\mathbf{1}[\Delta D_{i1} = -1]), \Delta D_{i1})}{Var(\Delta D_{i1})}. \tag{43}$$

Define

$$\begin{aligned} \omega &= 1 + \frac{Cov(\mathbf{1}[\Delta D_{i1} = -1]), \Delta D_{i1})}{Var(\Delta D_{i1})} \\ &= \frac{p^+(1-p^+) + p^+p^-}{p^+(1-p^+) + p^-(1-p^-) + 2p^+p^-}, \end{aligned} \tag{44}$$

where $p^+ = P(\Delta D_{i1} = 1)$ and $p^- = P(\Delta D_{i1} = -1)$. Note that $\omega \in [0, 1]$ and $\omega = 1/2$ when $1 - p^+ = p^-$. Thus,

$$\beta_1 = \widetilde{\beta}_1\omega + (\widetilde{\beta}_1 - \widetilde{\alpha}_1)(1 - \omega). \tag{45}$$

Combining equations (41), (42), and (45) completes the proof. □

Proof of Proposition 1 and Corollary

First suppose $P(\Delta D_{i1} = 0) > 0$. Under the impersistence and parallel trends assumptions,

$$\begin{aligned}
&E[\Delta Y_i \mid \Delta D_{i1} = 1] - E[\Delta Y_i \mid \Delta D_{i1} = 0] \\
&= \left(E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 1] - E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 0, D_{i00} = 1]\right) p \\
&\quad + \left(E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 1] - E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 0, D_{i10} = 1]\right) (1-p) \\
&= \big(E[Y_{i1}^{0\to1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1] - E[Y_{i1}^{0\to0} - Y_{i0}^{0} \mid \Delta D_{i1} = 0, D_{i00} = 1]\big) p \\
&\quad + \big(E[Y_{i1}^{0\to1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1] - E[Y_{i1}^{1\to1} - Y_{i0}^{1} \mid \Delta D_{i1} = 0, D_{i10} = 1]\big) (1-p) \\
&= E[Y_{i1}^{1} - Y_{i1}^{0} \mid \Delta D_{i1} = 1]p + E[Y_{i0}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1](1-p),
\end{aligned} \tag{46}$$

where $p = P(D_{i00} = 1 \mid \Delta D_{i1} = 0)$. Similarly,

$$\begin{aligned}
&E[\Delta Y_i \mid \Delta D_{i1} = 0] - E[\Delta Y_i \mid \Delta D_{i1} = -1] \\
&= \left(E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 0, D_{i00} = 1] - E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = -1]\right) p \\
&\quad + \left(E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = 0, D_{i10} = 1] - E[Y_{i1} - Y_{i0} \mid \Delta D_{i1} = -1]\right) (1-p) \\
&= \big(E[Y_{i1}^{0\to0} - Y_{i0}^{0} \mid \Delta D_{i1} = 0, D_{i00} = 1] - E[Y_{i1}^{1\to0} - Y_{i0}^{1} \mid \Delta D_{i1} = -1]\big) p \\
&\quad + \big(E[Y_{i1}^{1\to1} - Y_{i0}^{1} \mid \Delta D_{i1} = 0, D_{i10} = 1] - E[Y_{i1}^{1\to0} - Y_{i0}^{1} \mid \Delta D_{i1} = -1]\big) (1-p) \\
&= E[Y_{i0}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = -1]p + E[Y_{i1}^{1} - Y_{i1}^{0} \mid \Delta D_{i1} = -1](1-p).
\end{aligned} \tag{47}$$

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

$$\begin{aligned}
\beta_1 =& E[Y_{i1}^{1} - Y_{i1}^{0} \mid \Delta D_{i1} = 1]p\omega + E[Y_{i1}^{1} - Y_{i1}^{0} \mid \Delta D_{i1} = -1](1-p)(1-\omega) \\
&+ E[Y_{i0}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1](1-p)\omega + E[Y_{i0}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = -1]p(1-\omega).
\end{aligned} \tag{48}$$

Now consider the corollary case of $P(\Delta D_{i1} = 0) = 0$. From Lemma 1,

$$\begin{aligned}
\beta_1 =& (E[Y_{i1}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1] - E[Y_{i1}^{0} - Y_{i0}^{1} \mid \Delta D_{i1} = -1])/2 \\
=& (E[Y_{i1}^{0\to1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1] - E[Y_{i1}^{0\to1} - Y_{i0}^{1} \mid \Delta D_{i1} = 1])/2 \\
&- (E[Y_{i1}^{1\to0} - Y_{i0}^{1} \mid \Delta D_{i1} = -1] - E[Y_{i1}^{1\to1} - Y_{i0}^{1} \mid \Delta D_{i1} = -1])/2 \\
=& (E[Y_{i0}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1] + E[Y_{i1}^{1} - Y_{i1}^{0} \mid \Delta D_{i1} = -1])/2 \\
=& (E[Y_{i1}^{0\to1} - Y_{i0}^{0} \mid \Delta D_{i1} = 1] - E[Y_{i1}^{0\to0} - Y_{i0}^{0} \mid \Delta D_{i1} = 1])/2 \\
&- (E[Y_{i1}^{1\to0} - Y_{i0}^{1} \mid \Delta D_{i1} = -1] - E[Y_{i1}^{1\to0} - Y_{i0}^{0} \mid \Delta D_{i1} = -1])/2 \\
=& (E[Y_{i1}^{1} - Y_{i1}^{0} \mid \Delta D_{i1} = 1] + E[Y_{i0}^{1} - Y_{i0}^{0} \mid \Delta D_{i1} = -1])/2,
\end{aligned} \tag{49}$$

where the second and fourth lines again follow from impersistence and parallel trends. □

Proof of Proposition 2

The mover regression

$$\Delta Y_i = \tau + \sum_{j\neq 0} \beta_j (\mathbf{1}[\Delta D_{i1} = 1] - \mathbf{1}[\Delta D_{i1} = -1]) + \Delta \epsilon_i. \tag{50}$$

is nested by the model

$$\begin{aligned}
\Delta Y_i =& \widetilde{\tau} + \sum_{j\neq 0} \widetilde{\beta}_j (\mathbf{1}[\Delta D_{ij} = 1] - \mathbf{1}[\Delta D_{ij} = -1]) \qquad (51)\\
&+ \sum_{k} \sum_{j \neq 0,k} \widetilde{\delta}_{jk} \mathbf{1}[\Delta D_{ik} = 1]\mathbf{1}[\Delta D_{ij} = -1] + \Delta \widetilde{\epsilon}_i,\\
=& \widetilde{\tau} + \sum_{j\neq 0} \widetilde{\beta}_j D_{i00} D_{ij1} + \sum_{j\neq 0} (\widetilde{\delta}_{j0} - \widetilde{\beta}_j) D_{ij0} D_{i01} \qquad (52)\\
&+ \sum_{k} \sum_{j \neq 0,k} (\widetilde{\delta}_{jk} + \widetilde{\beta}_k - \widetilde{\beta}_j) D_{ij0} D_{ik1} + \Delta \widetilde{\epsilon}_i.
\end{aligned}$$

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

$$\widetilde{\beta}_j = E[\Delta Y_i \mid D_{i00} D_{ij1} = 1] - E[\Delta Y_i \mid D_{i\ell 0} D_{i\ell 1} = 1, \forall \ell], \tag{53}$$

$$\widetilde{\delta}_{j0} - \widetilde{\beta}_j = E[\Delta Y_i \mid D_{ij0} D_{i01} = 1] - E[\Delta Y_i \mid D_{i\ell 0} D_{i\ell 1} = 1, \forall \ell], \tag{54}$$

$$\widetilde{\delta}_{jk} + \widetilde{\beta}_k - \widetilde{\beta}_j = E[\Delta Y_i \mid D_{ij0} D_{ik1} = 1] - E[\Delta Y_i \mid D_{i\ell 0} D_{i\ell 1} = 1, \forall \ell]. \tag{55}$$

Under Assumption IO and the parallel trends condition, we can write

$$\begin{aligned}
\widetilde{\beta}_j =& (E[Y_{i1}^j - Y_{i0}^0 \mid D_{i00} D_{ij1} = 1] - E[Y_{i1}^0 - Y_{i0}^0 \mid D_{i00} D_{i01} = 1]) p_0\\
&+ (E[Y_{i1}^j - Y_{i0}^0 \mid D_{i00} D_{ij1} = 1] - E[Y_{i1}^j - Y_{i0}^j \mid D_{ij0} D_{ij1} = 1]) p_j\\
&+ \sum_{\ell \neq 0,j} (E[Y_{i1}^j - Y_{i0}^0 \mid D_{i00} D_{ij1} = 1] - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} D_{i\ell 1} = 1]) p_\ell\\
=& E[Y_{i1}^j - Y_{i1}^0 \mid D_{i00} D_{ij1} = 1] p_0 + E[Y_{i0}^j - Y_{i0}^0 \mid D_{i00} D_{ij1} = 1](1 - p_0) \qquad (56)\\
&+ \sum_{\ell \neq 0,j} (E[Y_{i1}^j - Y_{i0}^j \mid D_{ij0} D_{ij1} = 1] - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} D_{i\ell 1} = 1]) p_\ell,
\end{aligned}$$

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

$$\begin{aligned}
\widetilde{\delta}_{j0} =& \widetilde{\beta}_j - (E[Y_{i1}^j - Y_{i1}^0 \mid D_{ij0} D_{i01} = 1] p_0 + E[Y_{i0}^j - Y_{i0}^0 \mid D_{ij0} D_{i01} = 1](1 - p_0))\\
&- \sum_{\ell \neq 0,j} (E[Y_{i1}^j - Y_{i0}^j \mid D_{ij0} D_{ij1} = 1] - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0} D_{i\ell 1} = 1]) p_\ell\\
=& E[Y_{i1}^j - Y_{i1}^0 \mid D_{i00} D_{ij1} = 1] p_0 + E[Y_{i0}^j - Y_{i0}^0 \mid D_{i00} D_{ij1} = 1](1 - p_0) \qquad (57)\\
&- \Big(E[Y_{i1}^j - Y_{i1}^0 \mid D_{ij0} D_{i01} = 1] p_0 + E[Y_{i0}^j - Y_{i0}^0 \mid D_{ij0} D_{i01} = 1])(1 - p_0) \Big)
\end{aligned}$$

and

$$
\begin{align*}
\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_{i\ell 0}D_{i\ell 1} = 1, \forall \ell]) - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0}D_{i\ell 1} = 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_{i\ell 0}D_{i\ell 1} = 1, \forall \ell]) - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0}D_{i\ell 1} = 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 \tag{58} \\
&- 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_{i\ell 0}D_{i\ell 1} = 1, \forall \ell]) - E[Y_{i1}^\ell - Y_{i0}^\ell \mid D_{i\ell 0}D_{i\ell 1}, \forall \ell])p_\ell
\end{align*}
$$

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:

$$
\beta = \widetilde{\beta} + \sum_{k} \sum_{j \neq 0,k} \widetilde{\delta}_{jk} R_{jk}, \tag{59}
$$

where R_{jk} denotes the 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 (56)-(58) in to this expression shows that β will not in general identify a weighted average of causal parameters. □

Proof of Proposition 3 and Corollary

Let $\psi_i = P(\Delta D_i \neq 0)/P(\Delta D_i \neq 0 \mid X_i)$. For any two treatments c and d we have

$$\begin{aligned}
E[\Delta Y_i \rho_{it}^{c,d} \psi_i \mid X_i] &= (-1)^t E\left[\Delta Y_i D_{ic,1-t} \frac{D_{ict} E[D_{idt} D_{ic,1-t} \mid X_i] - D_{idt} E[D_{ict} D_{ic,1-t} \mid X_i]}{E[D_{ict} D_{ic,1-t} \mid X_i] E[D_{idt} D_{ic,1-t} \mid X_i]} \mid X_i\right] \\
&= (-1)^t \left(E\left[\Delta Y_i \mid D_{ic,1-t} D_{ict} = 1, X_i\right] - E\left[\Delta Y_i \mid D_{ic,1-t} D_{idt} = 1, X_i\right]\right) \\
&= E[Y_{i,1-t}^c - Y_{it}^c \mid D_{ic,1-t} D_{ict} = 1, X_i] - E[Y_{i,1-t}^c - Y_{it}^d \mid D_{ic,1-t} D_{idt} = 1, X_i] \\
&= E[Y_{it}^d - Y_{it}^c \mid D_{ic,1-t} D_{idt} = 1, X_i] \\
&= E[Y_{it}^d - Y_{it}^c \mid \Delta D_i \neq 0, X_i], \qquad (60)
\end{aligned}$$

provided $E[D_{ict} D_{ic,1-t} \mid X_i] E[D_{idt} D_{ic,1-t} \mid X_i] \neq 0$. Here the second-to-last line follows from Assumption CPT and, for $t = 0$, Assumption COI, while the last line follows by Assumption CEH. The same steps and assumptions show that

$$\begin{aligned}
E[\Delta Y_i (-\rho_{it}^{d,c}) \psi_i \mid X_i] &= E[Y_{it}^d - Y_{it}^c \mid D_{id,1-t} D_{ict} = 1, X_i] \\
&= E[Y_{it}^d - Y_{it}^c \mid \Delta D_i \neq 0, X_i]. \qquad (61)
\end{aligned}$$

Given the chain C_ℓ and set of constants w_m, we therefore have the main result.

$$\begin{aligned}
MATE_{jt} &= \int E[Y_{it}^j - Y_{it}^0 \mid \Delta D_i \neq 0, X_i] dP(X_i \mid \Delta D_i \neq 0) \\
&= \int \left(\sum_{m=1}^{M_{j\ell}} E[Y_{it}^{c_{j\ell m}} - Y_{it}^{c_{j\ell m-1}} \mid \Delta D_i \neq 0, X_i] \right) dP(X_i \mid \Delta D_i \neq 0) \\
&= E\left[\left(\sum_{m=1}^{M_{j\ell}} (w_m \Delta Y_i \rho_{it}^{c_{j\ell,m-1}, c_{j\ell m}} \psi_i + (1 - w_m) \Delta Y_i (-\rho_{it}^{c_{j\ell m}, c_{j\ell,m-1}}) \psi_i) \right) \frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)} \right] \\
&= E\left[\Delta Y_i \left(\sum_{m=1}^{M_{j\ell}} (w_m \rho_{it}^{c_{j\ell,m-1}, c_{j\ell m}} + (1 - w_m)(-\rho_{it}^{c_{j\ell m}, c_{j\ell,m-1}})) \right) \right]. \qquad (62)
\end{aligned}$$

The $J = 2$ corollary follows from the second-to-last lines of equations (60) and (61) by noting that

$$\begin{aligned}
MATE_{1t} = &\int E[Y_{it}^1 - Y_{it}^0 \mid \Delta D_{i1} \neq 0, X_i] dP(X_i \mid \Delta D_{i1} \neq 0) \\
= &\int E[\Delta Y_i \rho_{it}^{0,1} \psi_i \mid X_i] \frac{E[D_{i1t} D_{i0,1-t} \mid X_i]}{P(\Delta D_{i1} \neq 0 \mid X_i)} dP(X_i \mid \Delta D_{i1} \neq 0) \\
&+ \int E[\Delta Y_i (-\rho_{it}^{1,0}) \psi_i \mid X_i] \frac{E[D_{i0t} D_{i1,1-t} \mid X_i]}{P(\Delta D_{i1} \neq 0 \mid X_i)} dP(X_i \mid \Delta D_{i1} \neq 0) \\
= &E\left[\Delta Y_i \rho_{it}^{0,1} \psi_i \frac{E[D_{i1t} D_{i0,1-t} \mid X_i]}{P(\Delta D_{i1} \neq 0)} + \Delta Y_i (-\rho_{it}^{1,0}) \omega_i \frac{E[D_{i0t} D_{i1,1-t} \mid X_i]}{P(\Delta D_{i1} \neq 0)} \right] \\
= &E\left[\Delta Y_i \left(w_{it}^* \rho_{it}^{0,1} + (1 - w_{it}^*) \rho_{it}^{1,0} \right) \right]. \qquad \square \qquad (63)
\end{aligned}$$

Proof of Proposition 4

Let $\psi_i = P(\Delta D_i \neq 0)/P(\Delta D_i \neq 0 \mid X_i)$. For any two treatments c and d we have

$$\begin{aligned}
E[\Delta Y_i \kappa_i^{c,d} \psi_i \mid X_i] =& E\left[\Delta Y_i \frac{1}{2} \frac{D_{ic0}D_{id1}E[D_{id0}D_{ic1} \mid X_i] - D_{id0}D_{ic1}E[D_{ic0}D_{id1} \mid X_i]}{E[D_{id0}D_{ic1} \mid X_i]E[D_{ic0}D_{id1} \mid X_i]} \mid X_i\right] \\
=& \frac{1}{2}\left(E[\Delta Y_i \mid D_{ic0}D_{id1} = 1, X_i] - E[\Delta Y_i \mid D_{id0}D_{ic1} = 1, X_i]\right) \\
=& \frac{1}{2}\left(E[Y_{i1}^d - Y_{i0}^c \mid D_{ic0}D_{id1} = 1, X_i] - E[Y_{i1}^c - Y_{i0}^c \mid D_{ic0}D_{id1} = 1, X_i]\right) \\
& - \frac{1}{2}\left(E[Y_{i1}^c - Y_{i0}^d \mid D_{id0}D_{ic1} = 1, X_i] - E[Y_{i1}^c - Y_{i0}^c \mid D_{id0}D_{ic1} = 1, X_i]\right) \\
=& \frac{1}{2}\left(E[Y_{i1}^d - Y_{i1}^c \mid D_{ic0}D_{id1} = 1, X_i] + E[Y_{i0}^d - Y_{i0}^c \mid D_{id0}D_{ic1} = 1, X_i]\right) \\
=& \frac{1}{2}\left(E[Y_{i1}^d - Y_{i1}^c \mid \Delta D_i \neq 0, X_i] + E[Y_{i0}^d - Y_{i0}^c \mid \Delta D_i \neq 0, X_i]\right), \qquad (64)
\end{aligned}$$

provided $E[D_{id0}D_{ic1} \mid X_i]E[D_{ic0}D_{id1} \mid X_i] \neq 0$. Here the third equality follows from Assumptions CPT and COI, while the last line follows by Assumption CEH. Given the chain C_ℓ we thus have

$$\begin{aligned}
\frac{1}{2}(MATE_{j0} + MATE_{j1}) &= \int \frac{1}{2}\left(\sum_{t=0}^{1} E[Y_{it}^j - Y_{it}^0 \mid \Delta D_i \neq 0, X_i]\right) dP(X_i \mid \Delta D_i \neq 0) \\
&= \int \frac{1}{2}\left(\sum_{t=0}^{1}\sum_{m=1}^{M_{j\ell}} E[Y_{it}^{c_{j\ell m}} - Y_{it}^{c_{j\ell m-1}} \mid \Delta D_i \neq 0, X_i]\right) dP(X_i \mid \Delta D_i \neq 0) \\
&= E\left[\left(\sum_{m=1}^{M_{j\ell}} \Delta Y_i \kappa_i^{c_{j\ell,m-1},c_{j\ell m}} \psi_i\right) \frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)}\right] \\
&= E\left[\Delta Y_i \left(\sum_{m=1}^{M_{j\ell}} \kappa_i^{c_{j\ell,m-1},c_{j\ell m}}\right)\right], \qquad (65)
\end{aligned}$$

completing the proof. □

Extension to Many Time Periods

This appendix generalizes the MATE identification results to multi-period mover designs. Let $\Delta_{sr} V_i = V_{is} - V_{ir}$ denote the difference operator applied to variable V_{it} over periods $s > r$, and let ΔD_i collect all $\Delta_{sr} D_{ij}$. The multi-period version of Assumptions CPT and COI are

Assumption CPT′: For each treatment j, periods $t > s$, and x in the support of X_i,

$$E[Y_{it}^{\bar{j}_t \to j} - Y_{is}^{\bar{j}_s \to j} \mid D_{ijs} D_{ijt} = 1, X_i = x] = E[Y_{it}^{\bar{j}_t \to j} - Y_{is}^{\bar{j}_s \to j} \mid \Delta_{ts} D_{ij} = 1, X_i = x] \tag{66}$$

$$= E[Y_{it}^{\bar{j}_t \to j} - Y_{is}^{\bar{j}_s \to j} \mid \Delta_{ts} D_{ij} = -1, X_i = x], \tag{67}$$

where $\bar{j}_r$ denotes a vector of j's of length r.

Assumption COI′: For all treatments (j, k), periods $t > s$, and x in the support of X_i,

$$E[Y_{it}^{\bar{k}_t \to j} \mid D_{iks} D_{ijt} = 1, X_i = x] = E[Y_{i1}^{\bar{j}_t \to j} \mid D_{iks} D_{ijt} = 1, X_i = x], \tag{68}$$

where $\bar{k}_r$ denotes a vector of k's of length r.

Here Assumption CPT′ states that movers into or out of each treatment j between periods s and t would have, had they stayed at j, followed the same average outcome trends as each other and as treatment j stayers, conditional on the controls. Similarly Assumption COI′ states that movers into each treatment j from treatment k between periods s and t have the same time-t outcomes as if they had stayed at j. Assumption CEH from the text remains unmodified in the multi-period case.

We then have the following result, the proof of which follows by the same steps as that of Proposition 3:

Proposition 3′: Suppose Assumptions CPT′ and CEH hold, and for treatment j and periods $t > s$ there exists a chain $C_{j\ell}$ such that, for each $m = 1, \dots, M_{j\ell}$ and all x in the support of X_i either (i) $P(D_{i,m-1,s} D_{imt} = 1 \mid X_i = x) > 0$ and $P(D_{i,m-1,s} D_{i,m-1,t} = 1 \mid X_i = x) > 0$ or (ii) $P(D_{im,s} D_{i,m-1,t} = 1 \mid X_i = x) > 0$ and $P(D_{ims} D_{imt} = 1 \mid X_i = x) > 0$. Then

$$MATE_{jt} = E\left[\Delta_{ts} Y_i \left(\sum_{m=1}^{M_{j\ell}} (w_m \rho_{ist}^{c_{j\ell,m-1}, c_{j\ell m}} + (1 - w_m)(-\rho_{ist}^{c_{j\ell m}, c_{j\ell,m-1}}))\right)\right], \tag{69}$$

where the w_m are constants such that $w_m = 0$ if (i) fails for m and $w_m = 1$ if (ii) fails for m, and where

$$\tilde{\rho}_{it}^{c,d} = (-1)^{\mathbf{1}[t>s]} D_{ic,s} \frac{D_{ict} E[D_{idt} D_{ics} \mid X_i] - D_{idt} E[D_{ict} D_{ics} \mid X_i]}{E[D_{ict} D_{ics} \mid X_i] E[D_{idt} D_{ics} \mid X_i]} \frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)}. \tag{70}$$

If moreover Assumption COI holds, equations (27) and (28) also hold for periods $t < s$.

We also have the following generalization of Proposition 4, the proof of which follows similarly

Proposition 4′: Suppose Assumptions CPT′, CEH, and COI′ hold, and for treatment j and periods (t, s) there exists a chain $C_{j\ell}$ such that, for each $m = 1, \ldots, M_{j\ell}$ and all x in the support of X_i, $P(D_{i,m-1,s}D_{imt} = 1 \mid X_i = x) > 0$ and $P(D_{ims}D_{i,m-1,t} = 1 \mid X_i = x) > 0$. Then

$$\frac{1}{2}(MATE_{js} + MATE_{jt}) = E\left[\Delta Y_i \left(\sum_{m=1}^{M_{j\ell}} \kappa_i^{c_{j\ell,m-1}, c_{j\ell m}}\right)\right], \tag{71}$$

where

$$\tilde{\kappa}_i^{c,d} = \frac{1}{2}\frac{D_{ics}D_{idt}E[D_{ids}D_{ict} \mid X_i] - D_{ids}D_{ict}E[D_{ics}D_{idt} \mid X_i]}{E[D_{ids}D_{ict} \mid X_i]E[D_{ics}D_{idt} \mid X_i]}\frac{P(\Delta D_i \neq 0 \mid X_i)}{P(\Delta D_i \neq 0)}. \tag{72}$$

Analogous results for the specification tests and efficient estimators derived in Section 3.3 would again apply here, with overidentification resulting from either many chains or many time period pairs satisfying Propositions 3′ and 4′.