Difference-in-Differences Estimates under Selective Migration

Roberto Valli*

February 23, 2026

Abstract

Difference-in-differences (DiD) designs are widely used to estimate causal effects of location-based treatments on individual outcomes. Yet when treatments trigger selective migration, observed changes in location-level outcomes conflate behavioral responses of stayers with compositional shifts caused by selective entry and exit. This paper develops a formal decomposition of DiD estimands in the presence of treatment-induced migration. Building on principal stratification, I show that the aggregate DiD estimand decomposes into a behavioral component and compositional terms driven by selective exits and entries. The within-unit DiD with individual fixed effects identifies the Survivor Average Treatment Effect (SATE) for stayers, but not the average treatment effect on the treated (ATT) for the full pre-treatment population. I characterize the ATT–SATE gap and propose three strategies for applied researchers to quantify compositional effects. Extensions address migration across treatment and control areas, where contamination and depletion biases compound the compositional effect, and staggered treatment adoption, where migration bias accumulates with exposure time.

Keywords: Difference-in-differences, migration, panel data, policy evaluation, bias correction.
Word count: 3'345 without references.

* Max Weber Fellow, European University Institute. 50014 Fiesole, Italy. Email: `roberto.valli@eui.eu`

Introduction

Difference-in-differences (DiD) is among the most widely used research designs in the social sciences, and a large methodological literature has refined the design to handle staggered adoption, heterogeneous treatment effects, and violations of parallel trends (Abadie et al., 2025; Roth et al., 2023). A fundamental challenge arises, however, when the treatment itself changes the composition of the observed population. Many place-based treatments, such as policy changes and environmental disasters, trigger migration across locations. Individuals selectively enter or exit the treated area, and the population observed after treatment differs systematically from the one observed before. When this migration is correlated with potential outcomes, aggregate DiD estimates conflate the *behavioral response* of individuals who remain with the *compositional shift* caused by selective turnover.

Applied researchers typically pursue one of two strategies, each with its limitations. First, they estimate aggregate DiD at the locality level, which captures the total effect including composition changes but cannot separate behavioral from compositional mechanisms. Second, they estimate within-individual models with individual fixed effects restricted to "stayers," which absorb time-invariant heterogeneity but condition on a post-treatment variable—the decision to stay—and identify the causal effect only for a selected subpopulation. The first strategy ignores a potentially serious violation of SUTVA assumptions, whereas the second targets a causal quantity that might deviate from the average treatment effect on the treated (ATT).

This paper develops a formal decomposition of the DiD estimand in panel settings with treatment-induced migration. I make three contributions. First, I decompose the aggregate DiD into a behavioral component (the Survivor Average Treatment Effect, or SATE) and compositional terms driven by the departure of escapees and arrival of followers (Proposition 2). Second, I show that the within-unit DiD with individual fixed effects identifies the SATE, and characterize when and how it diverges from the ATT through a selection parameter δ (Proposition 3). Third, I propose three practical estimation strategies—direct compositional decomposition, Lee-type bounds, and a calibrated sensitivity analysis—that allow researchers to quantify compositional effects and assess the ATT–SATE gap under different data availability scenarios. Two extensions address settings where migration crosses the treatment–control boundary (creating contamination and depletion biases) and where treatment is adopted in a staggered rollout (where migration bias accumulates with exposure time).

The approach builds on the principal stratification framework of Marbach (2024, 2025) and relates to recent work on DiD with compositional changes. Sant'Anna and Xu (2025) derive efficient estimators under selection on observables for repeated cross-sections. Rathnayakea et al. (2024) construct bounds under endogenous selection. Ghanem et al. (2024) extend changes-in-changes under rank invariance. Lee (2009) proposes monotonicity-based bounds for nonrandom attrition. This paper differs in providing ways to use available individual-level information to gauge compositional effects. Because escapees are observed at $t = 0$ before they leave, their pre-treatment characteristics are directly measurable without distributional assumptions or the requirement that all selection determinants be observed. This enables point estimation of compositional effects and transparent calibration of the ATT–SATE gap through a single parameter.

A parallel literature studies DiD under interference, where treatment changes outcomes for untreated units. Mealli and Viviens (2025) show that, under unknown interference, the standard DiD identifies a *difference* of two causal effects—the total effect on the treated minus the average spillover on the control—rather than a single treatment effect. Xu (2025) develops doubly robust estimators for the direct average treatment effect using exposure mappings, and Butts (2024) proposes ring-based estimators that exploit spatial distance to separate direct from spillover effects. These papers address *behavioral* spillovers, meaning that treatment in one area changes outcomes in another. Instead, the present paper addresses a complementary *compositional* channel, where treatment changes who is observed in each area, even absent any behavioral spillover. The two mechanisms can operate simultaneously, and the interference extension in Section shows how compositional and spillover biases compound.

Framework

Consider the canonical setting with agents $i \in \{1, \dots, N\}$ observed over two periods $t \in \{0, 1\}$, located in localities partitioned into treated ($\mathcal{T}$) and control ($\mathcal{C}$) areas. Treatment $D_j \in \{0, 1\}$ is assigned at the locality level. Let $Y_{it}(d)$ denote potential outcomes and $S_i(d) \in \{0, 1\}$ the staying indicator, where $S_i(d) = 1$ means agent i remains in locality j under treatment status d.

In a first simplified setting, let us also define two assumptions:

Assumption 1 (No Interference)**.** *Treatment in $\mathcal{T}$ does not affect outcomes or migration decisions in $\mathcal{C}$: for all agents i initially in a control locality, $Y_{it}(d) = Y_{it}(0)$ and $S_i(d) = S_i(0)$ for all d.*

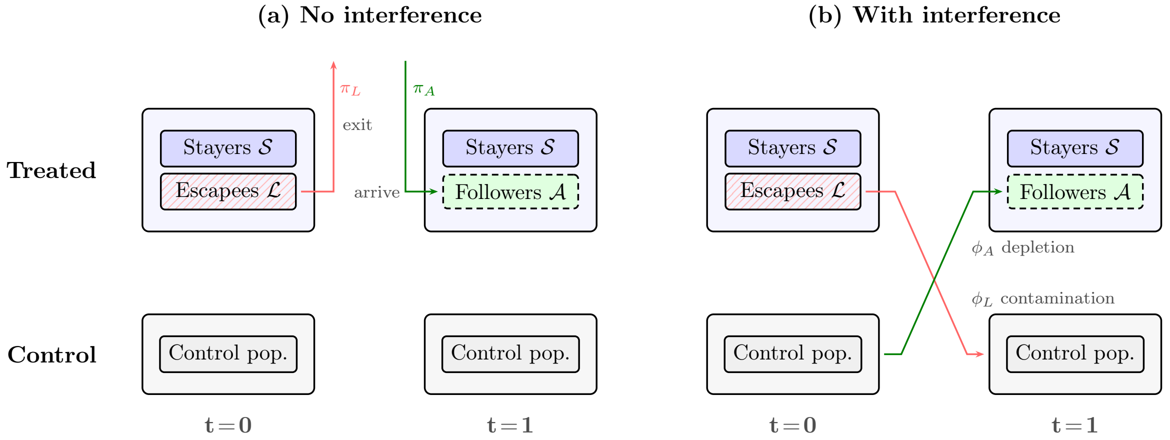
This assumption states that treated and control areas are behaviorally disconnected over the study window. A policy shock in treated places may change who stays and how outcomes evolve there, but it cannot directly alter outcomes or migration choices of agents initially in control places. This isolates the compositional channel within treated areas and rules out contamination of the control trend through spillovers.

Assumption 2 (No Anticipation)**.** $Y_{i0}(1) = Y_{i0}(0)$ *for all* i*.*

Intuitively, no anticipation requires that agents do not adjust outcomes before treatment is active. Baseline outcomes at $t = 0$ are therefore clean pre-treatment measurements, rather than already incorporating behavioral responses to expected future treatment. This is what allows pre-treatment differences to be interpreted as compositional characteristics rather than early treatment effects.

Following Marbach (2024), each agent in $\mathcal{T}$ belongs to one of four principal strata defined by $(S_i(0), S_i(1))$: **stayers** ($\mathcal{S}$: remain regardless), **escapees** ($\mathcal{L}$: leave because of treatment), **followers** ($\mathcal{A}$: arrive because of treatment), and **never-present** ($\mathcal{N}$: absent regardless). At $t = 0$, the treated population comprises $\mathcal{S} \cup \mathcal{L}$; at $t = 1$, it comprises $\mathcal{S} \cup \mathcal{A}$. Let $\pi_S, \pi_L, \pi_A, \pi_N$ denote the population shares of each stratum among all units connected to treated localities, with $\pi_S + \pi_L + \pi_A + \pi_N = 1$. The stayer share in the pre-treatment observed population is $w_0 = \pi_S/(\pi_S + \pi_L)$, and in the post-treatment population $w_1 = \pi_S/(\pi_S + \pi_A)$. Figure 1 illustrates the observability structure.

Figure 1: Principal strata



(a) No interference

(b) With interference

Note: At $t = 0$ the treated area contains stayers and future escapees; at $t = 1$ it contains stayers and followers. (a) Without interference: escapees exit and followers enter from outside the study; the control group is unaffected. (b) With interference: a fraction ϕ_L of escapees relocate to control localities (contamination) and a fraction ϕ_A of followers originate from control localities (depletion), creating crossing flows that bias the control-group counterfactual.

I define three causal estimands. The ATT for the pre-treatment population is:

$$\tau^{ATT} = \mathbb{E}\big[Y_{i1}(1) - Y_{i1}(0) \mid D_j = 1, i \in \mathcal{S} \cup \mathcal{L}\big] \tag{1}$$

The Survivor Average Treatment Effect (SATE) is:

$$\tau^{SATE} = \mathbb{E}\big[Y_{i1}(1) - Y_{i1}(0) \mid D_j = 1,\ i \in \mathcal{S}\big] \tag{2}$$

The Escapee Average Treatment Effect (EATE) is:

$$\tau^{EATE} = \mathbb{E}\big[Y_{i1}(1) - Y_{i1}(0) \mid D_j = 1,\ i \in \mathcal{L}\big] \tag{3}$$

This quantity is hypothetical: it requires the potential outcome $Y_{i1}(1)$ for units who leave under treatment. The three estimands are linked by:

$$\tau^{ATT} = \frac{\pi_S}{\pi_S + \pi_L} \cdot \tau^{SATE} + \frac{\pi_L}{\pi_S + \pi_L} \cdot \tau^{EATE} \tag{4}$$

Define the *selection parameter* $\delta \equiv \tau^{EATE} - \tau^{SATE}$, so that:

$$\tau^{ATT} = \tau^{SATE} + \frac{\pi_L}{\pi_S + \pi_L} \cdot \delta \tag{5}$$

When $\delta = 0$, the SATE equals the ATT. When $\delta \neq 0$, the within-unit DiD over- or under-estimates the ATT depending on the sign and magnitude of δ.

Decomposition Results

Let $\bar{Y}_{j,t}$ denote the mean outcome among agents observed at locality j at time t. The pre-treatment mean in a treated locality is $\bar{Y}_{j,0} = w_0\bar{Y}_{\mathcal{S},0} + (1-w_0)\bar{Y}_{\mathcal{L},0}$, where $w_0 = N_{\mathcal{S}}/(N_{\mathcal{S}}+N_{\mathcal{L}})$ is the stayer share. The post-treatment mean is $\bar{Y}_{j,1} = w_1\bar{Y}_{\mathcal{S},1}(1) + (1-w_1)\bar{Y}_{\mathcal{A},1}(1)$, where $w_1 = N_{\mathcal{S}}/(N_{\mathcal{S}} + N_{\mathcal{A}})$.

Proposition 1 (Treated First-Difference Decomposition)**.** *The first-difference for treated locality j decomposes as:*

$$\bar{Y}_{j,1} - \bar{Y}_{j,0} = \underbrace{\big(\bar{Y}_{\mathcal{S},1}(1) - \bar{Y}_{\mathcal{S},0}\big)}_{\textit{Stayer change}} + \underbrace{(1-w_0)\big(\bar{Y}_{\mathcal{S},0} - \bar{Y}_{\mathcal{L},0}\big)}_{\textit{Leaver composition}} + \underbrace{(1-w_1)\big(\bar{Y}_{\mathcal{A},1}(1) - \bar{Y}_{\mathcal{S},1}(1)\big)}_{\textit{Follower composition}} \tag{6}$$

The leaver term is positive when escapees had lower outcomes than stayers at baseline, meaning that removing low-Y units raises the average. The follower term is positive when followers have higher outcomes than stayers post-treatment. The control group decomposes analogously (Appendix A1).

Assumption 3 (Parallel Trends at the Aggregate Level)**.** $\mathbb{E}\big[\bar{Y}_{j,1}(0) - \bar{Y}_{j,0} \mid j \in \mathcal{T}\big] = \mathbb{E}\big[\bar{Y}_{j,1} - \bar{Y}_{j,0} \mid j \in \mathcal{C}\big]$

Assumption 4 (Parallel Trends for Stayers)**.** $\mathbb{E}\big[Y_{i1}(0) - Y_{i0}(0) \mid i \in \mathcal{S},\, j \in \mathcal{T}\big] = \mathbb{E}\big[Y_{i1}(0) - Y_{i0}(0) \mid j \in \mathcal{C}\big]$

Assumption 4 strengthens Assumption 3 by requiring parallel trends specifically for stayers, which is needed to isolate the behavioral component as the SATE.

Proposition 2 (Aggregate DiD Decomposition)**.** *Under Assumptions 1–4, the aggregate DiD estimand decomposes as:*

$$\theta^{agg} = \underbrace{\tau^{SATE}}_{Behavioral} + \underbrace{\theta^{comp}}_{Compositional} \tag{7}$$

where:

$$\theta^{comp} = (1 - w_0)\big(\bar{Y}_{\mathcal{S},0} - \bar{Y}_{\mathcal{L},0}\big) + (1 - w_1)\big(\bar{Y}_{\mathcal{A},1}(1) - \bar{Y}_{\mathcal{S},1}(1)\big) - \theta^{comp,C} \tag{8}$$

and $\theta^{comp,C}$ denotes the analogous compositional terms in the control group, capturing natural turnover.

Under Assumptions 1–4, the DiD removes the common time trend and natural compositional shifts. What remains is the behavioral effect for stayers plus the compositional distortion from treatment-induced exits and entries beyond natural turnover.

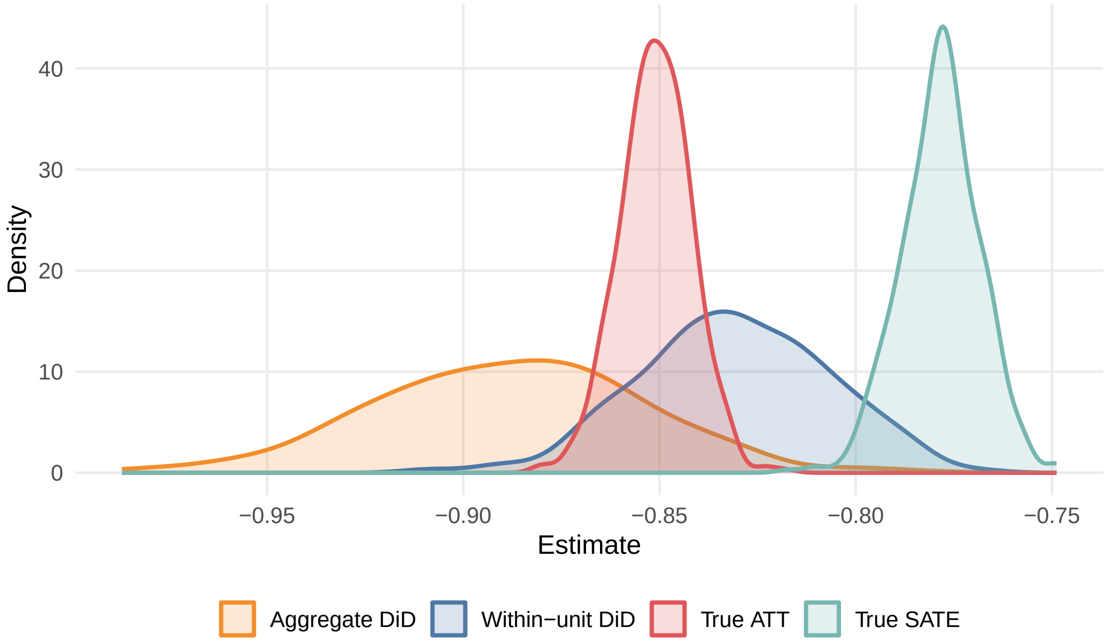
Proposition 3 (Within-Unit DiD Identifies the SATE)**.** *Under Assumptions 1, 2, and 4, the within-unit DiD coefficient θ^{ind} from $Y_{ijt} = \rho_i + \gamma_t + \theta^{ind} D_{jt} + \varepsilon_{ijt}$, estimated on stayers (where individual fixed effects absorb locality effects since stayers do not change locality), identifies:*

$$\theta^{ind} = \tau^{SATE} \tag{9}$$

By restricting to units observed in both periods, the within-unit DiD eliminates compositional effects. Individual fixed effects absorb time-invariant heterogeneity, and under parallel trends for stayers, the DiD coefficient recovers the SATE. A useful implication is that $\theta^{agg} - \theta^{ind} = \theta^{comp}$, which serves as a diagnostic for compositional change when the two estimators use commensurable samples and weighting schemes. Proofs are in Appendix A1.

In practice, most DiD studies report only $\hat{\theta}^{agg}$. When treatment triggers selective migration, this estimate includes compositional change that is not a behavioral response—the sign or magnitude of the "effect" may be driven by who leaves or arrives, not by how stayers respond. The within-unit DiD, which many papers report as a robustness check, identifies a valid causal effect but for a different population than the ATT. The gap between the two estimators is informative: a large $|\hat{\theta}^{agg} - \hat{\theta}^{ind}|$ signals that compositional change is substantial, and the strategies developed below become essential. Figure 2 illustrates these relationships using Monte Carlo simulations (see Appendix A3).

Figure 2: Distribution of estimands across 500 Monte Carlo replications



Note: (N = 1,200, J = 60, π_L = 0.25, π_A = 0.20, $\delta = -0.6$). The aggregate DiD (orange) is shifted away from the true ATT (red) by the compositional effect. The within-unit DiD (blue) tracks the true SATE (dashed). The ATT lies between the two estimators, but closer to the SATE when the escapee share is moderate.

Identification and Estimation

The decomposition separates the aggregate DiD into behavioral and compositional components. Two questions remain: how large is the compositional effect, and how far is the SATE from the ATT? I propose three strategies to answer these questions, arranged from most to least data-demanding. Table 1 summarizes the assumptions required by each strategy.

Table 1: Assumptions required by each identification strategy

	Assumption Required?					
Strategy	No Interf. (A1)	No Antic. (A2)	PT Aggregate (A3)	PT Stayers (A4)	Mono-tonicity (A5)	Individual Tracking Data
Aggregate DiD ($\hat{\theta}^{agg}$)	✓	✓	✓			
Within-unit DiD ($\hat{\tau}^{SATE}$)	✓	✓		✓		✓
Direct decomposition	✓	✓	✓	✓		✓
Lee bounds	✓	✓		✓	✓	
Sensitivity analysis	✓	✓		✓		✓

Strategy 1: Direct decomposition

When individual-level panel data with persistent identifiers is available, agents can be classified into stayers, escapees, and followers. For instance, researchers working with population registries might be able to track individuals over time and across localities. The compositional effect $\hat{\theta}^{comp}$ is estimated from four observable quantities: the conditional shares $(1-\hat{w}_0)$ and $(1-\hat{w}_1)$, the pre-treatment stayer–escapee gap $\hat{\Delta}_{SL} = \bar{Y}_{\mathcal{S},0} - \bar{Y}_{\mathcal{L},0}$, and the post-treatment follower–stayer gap $\hat{\Delta}_{AS} = \bar{Y}_{\mathcal{A},1} - \bar{Y}_{\mathcal{S},1}$. Based on this information, the estimated compositional effect is:

$$\hat{\theta}^{comp} = (1-\hat{w}_0)\cdot\hat{\Delta}_{SL} + (1-\hat{w}_1)\cdot\hat{\Delta}_{AS} - \hat{\theta}^{comp,C} \quad (10)$$

where $\hat{\theta}^{comp,C}$ captures natural turnover in the control group. An internal validity check is whether $\hat{\theta}^{comp} \approx \hat{\theta}^{agg} - \hat{\theta}^{ind}$. In MC simulations, this strategy is able to closely track true compositional effects θ^{comp} across levels of migration intensity (see Figure A2).

Strategy 2: Lee-type bounds

When individual tracking is unavailable, Lee-type bounds provide nonparametric bounds on the SATE. The core idea is to trim the excess attrition caused by treatment from the outcome distribution (Lee, 2009). The bounds can be computed under more or less demanding monotonicity assumptions.

Assumption 5 (Monotonicity of Migration)**.** *Treatment weakly increases the probability of leaving: $S_i(1) \leq S_i(0)$ for all i (no followers).*

When only exits are treatment-induced, Assumption 5 applies and the researcher computes the excess attrition rate $\hat{q}_L = \hat{P}(\text{leave} \mid \mathcal{T}) - \hat{P}(\text{leave} \mid \mathcal{C})$, then trims the $\hat{q}_L$ fraction of pre-treatment treated observations from each tail of the outcome distribution. Running the DiD on each trimmed sample yields bounds $\hat{\theta}^{LB} \leq \tau^{SATE} \leq \hat{\theta}^{UB}$. The trimming is a bounding construction: it identifies the range of estimates consistent with worst-case selection under monotonicity. *When both exits and entries are treatment-induced,* the procedure extends by defining separate excess exit and entry rates, $\hat{q}_L$ and $\hat{q}_A$, and trimming $\hat{q}_L$ from the pre-treatment sample and $\hat{q}_A$ from the post-treatment sample. This requires a weaker assumption that treatment weakly increases exits *and* weakly increases entries, each relative to the control group. The four tail combinations yield the bounds described in Appendix A2. Standard errors are obtained via nonparametric bootstrap.

Strategy 3: Sensitivity analysis for the ATT

The within-unit DiD identifies the SATE, but researchers typically are interested in the ATT. Rather than assuming $\delta = 0$, a sensitivity analysis makes the dependence on δ transparent:

$$\tau^{ATT}(\delta) = \hat{\tau}^{SATE} + \frac{\hat{\pi}_L}{\hat{\pi}_S + \hat{\pi}_L} \cdot \delta \tag{11}$$

This is linear in δ, with slope equal to the escapee share, as demonstrated by the MC-based sensitivity plot in Figure A1. A sensitivity plot displays $\tau^{ATT}(\delta)$ against δ, with vertical markers at calibration points indicating the plausible range of δ values based on the data, and a horizontal band at the sign-reversal threshold. The sign-reversal threshold is $\delta^* = -\hat{\tau}^{SATE} \cdot (\hat{\pi}_S + \hat{\pi}_L)/\hat{\pi}_L$, the value at which $\tau^{ATT}(\delta) = 0$. The researcher can assess whether the values of δ that would overturn qualitative conclusions fall within the plausible range.

Scholars can use several calibration strategies to anchor plausible values of δ. The pre-treatment stayer–escapee gap $\hat{\Delta}_{SL}$ provides a direct measure of selection on levels: a natural benchmark is $|\delta| \leq \kappa \cdot |\hat{\Delta}_{SL}|$, namely that selection on treatment effects be bounded by a multiple of the observable pre-treatment gap. Figure A4 validates that $\hat{\Delta}_{SL}$ responds monotonically to the intensity of selection on levels and demonstrates how it anchors the sensitivity plot at $\kappa \in \{0.5, 1, 2\}$. Alternatively, if the analysis reports treatment effect heterogeneity by subgroups, the range of subgroup-specific SATEs can be used to bound $|\delta|$.

Extensions

The baseline framework assumes that treatment-induced migration does not cross the treatment–control boundary and that treatment is adopted simultaneously. This section summarizes two extensions, developed fully in Appendices A4 and A5, that relax each restriction in turn.

Migration across treatment and control areas

When escapees relocate to control localities or followers are drawn from them, the no-interference assumption (Assumption 1) fails. Define the *contamination rate* ϕ_L as the share of escapees who relocate to $\mathcal{C}$, and the *depletion rate* ϕ_A as the share of followers originating from $\mathcal{C}$. Appendix A4 shows that the aggregate DiD under interference decomposes as:

$$\theta^{agg,*} = \tau^{SATE} + \theta^{comp} - \text{Bias}_L - \text{Bias}_A \tag{12}$$

where Bias_L and Bias_A capture contamination and depletion of the control group, respectively. These biases are additive: each depends on the migrant share in the control population and the outcome gap between migrants and native control agents. The within-unit DiD remains approximately unbiased for the SATE when the migrant share in the control group is small. When the direction of selection is known, sign restrictions yield one-sided bounds on the no-interference estimand. When migration is spatially concentrated, "donut" control groups that exclude localities near the treatment boundary provide a contamination-free comparison, subject to a parallel-trends condition for the donut sample.

These results have direct implications for the three estimation strategies developed in Section . Direct decomposition extends when migrant destinations are observed: Bias_L is estimable from the migrant share and the outcome gap between relocated escapees and control natives. Lee-type bounds should be computed on the donut control group to avoid contamination. The sensitivity analysis in δ remains valid as a bound under sign restrictions on contamination, but becomes a two-parameter problem (δ, ϕ_L) if the direction of contamination is unknown.

Staggered adoption

Many place-based policies are rolled out in stages, creating staggered treatment adoption. Appendix A5 extends the two-period decomposition to this setting. The key insight is that the decomposition in Proposition 2 applies cell-by-cell to each cohort-event pair (g, e), so migration bias accumulates with exposure time. The rate of accumulation depends on the migration process. Under *contained one-move* dynamics ($m = 1$), where each agent moves at most once, the selection-bias component plateaus after the first period of exposure. Under *cascade* dynamics ($m > 1$), where agents may move repeatedly, selection bias grows with exposure, so that earlier-treated cohorts carry more compositional distortion by later calendar periods. This creates heterogeneity in bias across cohort-event cells that is observationally similar to treatment effect heterogeneity.

Clean staggered estimators, such as those proposed by Borusyak, Jaravel, and Spiess (2024), Callaway and Sant'Anna (2021), and Sun and Abraham (2021), eliminate bias from heterogeneous treatment effects and negative weighting, but do not address the compositional channel. Under migration, these estimators identify cohort-specific $SATE(g,t)$ plus a compositional term, not $ATT(g,t)$. The decomposition and sensitivity tools developed in this paper apply to each (g,e) cell, and the calibrated sensitivity analysis extends naturally to the staggered case by allowing δ and Δ_{SL} to vary by cohort.

Discussion

This paper formalizes a problem that applied researchers often encounter informally: what quantity do aggregate and individual-level DiD estimates retrieve in settings with endogenous migration? The key innovation is leveraging the panel structure to observe escapees before they leave, enabling point estimation of compositional effects without distributional assumptions, something unavailable in cross-sectional principal stratification (Marbach, 2024), and underexplored by methods requiring selection on observables (Sant'Anna & Xu, 2025). Relative to the growing literature on DiD under interference (Butts, 2024; Mealli & Viviens, 2025; Xu, 2025), the compositional channel arises even when Assumption 1 holds perfectly, because who is *observed* changes even if no outcomes are affected across treatment areas.

Based on the framework, I recommend that applied researchers in DiD settings with potential treatment-induced migration: (i) report both aggregate and within-unit DiD estimates, as their gap is a first-order diagnostic for compositional effects; (ii) characterize selection by documenting leaving and arrival rates and comparing pre-treatment outcomes of stayers and leavers; (iii) estimate the compositional effect directly when tracking data is available; and (iv) conduct a sensitivity analysis for the ATT using calibrated values of δ.

The final section of the paper extends the framework in two directions. First, to settings where migration crosses the treatment–control boundary, creating contamination and depletion biases (Appendix A4). Second, to staggered treatment adoption settings, where migration bias accumulates with exposure time (Appendix A5). The interference extension is particularly relevant for place-based designs with geographically proximate control groups. The staggered extension shows that clean estimators addressing heterogeneous effects do not resolve the compositional channel.

References

Abadie, A., Angrist, J., Frandsen, B., & Pishke, J.-S. (2025). *Harvesting Differences-in-Differences and Event-Study Evidence* (Working Paper No. 34550). National Bureau of Economic Research. https://doi.org/10.3386/w34550

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event study designs: Robust and efficient estimation. *Review of Economic Studies*, *91*(6), 3253–3286.

Butts, K. (2024). *Difference-in-Differences with Spatial Spillovers* [Working paper, University of Colorado Boulder].

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods [Themed Issue: Treatment Effect 1]. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/https://doi.org/10.1016/j.jeconom.2020.12.001

Ghanem, D., Hirshleifer, S., Kédagni, D., & Ortiz-Becerra, K. (2024). Correcting Attrition Bias using Changes-in-Changes. https://arxiv.org/abs/2203.12740

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies*, *76*(3), 1071–1102. https://doi.org/10.1111/j.1467-937X.2009.00536.x

Marbach, M. (2024). Causal Effects, Migration, and Legacy Studies. *American Journal of Political Science*, *68*(4), 1447–1459. https://doi.org/https://doi.org/10.1111/ajps.12809

Marbach, M. (2025). Compositional Effects, Internal Migration and Electoral Outcomes. https://osf.io/preprints/socarxiv/pq3bd_v1

Mealli, F., & Viviens, J. (2025). *Difference-in-Differences in the Presence of Unknown Interference* [arXiv:2512.21176v2].

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics*, *37*(2), 187–204. https://doi.org/10.1080/07350015.2016.1227711

Rathnayakea, G., Negia, A., Bartalottia, O., & Zhao, X. (2024). Difference-in-Differences with Sample Selection.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, *235*(2), 2218–2244. https://doi.org/https://doi.org/10.1016/j.jeconom.2023.03.008

Sant'Anna, P. H. C., & Xu, Q. (2025). Difference-in-differences with Compositional Changes. *Journal of Econometrics*, *0*(0), 101–122. https://doi.org/https://doi.org/10.1016/j.jeconom.2020.06.003

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects [Themed Issue: Treatment Effect 1]. *Journal of Econometrics*, *225*(2), 175–199. https://doi.org/https://doi.org/10.1016/j.jeconom.2020.09.006

Xu, R. (2025). *Difference-in-Differences with Interference* [arXiv:2306.12003v6].

Supplementary Materials

A1 Proofs of results in main text

A1.1 Proof of Proposition 1 (Treated First-Difference Decomposition)

Rewrite the pre-treatment mean by adding and subtracting $\bar{Y}_{\mathcal{S},0}$:

$$\bar{Y}_{j,0} = w_0\,\bar{Y}_{\mathcal{S},0} + (1-w_0)\,\bar{Y}_{\mathcal{L},0} = \bar{Y}_{\mathcal{S},0} - (1-w_0)\big(\bar{Y}_{\mathcal{S},0} - \bar{Y}_{\mathcal{L},0}\big) \tag{13}$$

Similarly, rewrite the post-treatment mean:

$$\bar{Y}_{j,1} = w_1\,\bar{Y}_{\mathcal{S},1}(1) + (1-w_1)\,\bar{Y}_{\mathcal{A},1}(1) = \bar{Y}_{\mathcal{S},1}(1) + (1-w_1)\big(\bar{Y}_{\mathcal{A},1}(1) - \bar{Y}_{\mathcal{S},1}(1)\big) \tag{14}$$

Subtracting yields the result. □

A1.2 Proof of Proposition 2 (Aggregate DiD Decomposition)

Apply Proposition 1 to the treated group and the analogous decomposition to the control group. The aggregate DiD is $\theta^{agg} = \mathbb{E}[\Delta_j^T] - \mathbb{E}[\Delta_j^C]$.

For the treated group, the stayer change is $\bar{Y}_{\mathcal{S},1}(1) - \bar{Y}_{\mathcal{S},0}$. By Assumption 2, $\bar{Y}_{\mathcal{S},0} = \bar{Y}_{\mathcal{S},0}(0)$, so:

$$\bar{Y}_{\mathcal{S},1}(1) - \bar{Y}_{\mathcal{S},0}(0) = \tau^{SATE} + \mathbb{E}\big[Y_{i1}(0) - Y_{i0}(0) \mid i \in \mathcal{S},\, j \in \mathcal{T}\big] \tag{15}$$

Under Assumption 4, the counterfactual trends cancel and the DiD differences out the trend, leaving $\theta^{agg} = \tau^{SATE} + \theta^{comp}$, where θ^{comp} collects the treated compositional terms net of the control group's natural turnover. □

A1.3 Proof of Proposition 3 (Within-Unit DiD Identifies the SATE)

The within-unit DiD for stayer i in treated locality j is $Y_{i1}(1) - Y_{i0}(0)$. Differencing against the mean control-group change $\mathbb{E}[Y_{i1}(0) - Y_{i0}(0) \mid j \in \mathcal{C}]$ and taking expectations over stayers:

$$\begin{aligned}\theta^{ind} &= \mathbb{E}\big[Y_{i1}(1) - Y_{i0}(0) \mid i \in \mathcal{S}, j \in \mathcal{T}\big] - \mathbb{E}\big[Y_{i1}(0) - Y_{i0}(0) \mid j \in \mathcal{C}\big] \\ &= \tau^{SATE} + \underbrace{\mathbb{E}\big[Y_{i1}(0) - Y_{i0}(0) \mid i \in \mathcal{S},\, j \in \mathcal{T}\big] - \mathbb{E}\big[Y_{i1}(0) - Y_{i0}(0) \mid j \in \mathcal{C}\big]}_{=0 \text{ by Assumption } 4}\end{aligned} \tag{16}$$

□

A1.4 Proof of Proposition 4 (Contamination Bias)

The no-interference aggregate DiD uses $\bar{Y}_{C,1}^{nat}$ as the post-treatment control mean. Under contamination, the observed control mean is $\bar{Y}_{C,1}^{*}$ from the weighted average in Appendix A4. The contaminated aggregate DiD is:

$$\theta^{agg,*} = \big(\bar{Y}_{j,1} - \bar{Y}_{j,0}\big) - \big(\bar{Y}_{C,1}^{*} - \bar{Y}_{C,0}\big) \tag{17}$$

Substituting $\bar{Y}_{C,1}^{*} = \bar{Y}_{C,1}^{nat} + \frac{\phi_L N_{\mathcal{L}}}{N_{C,1}^{*}}(\bar{Y}_{\mathcal{L}\to\mathcal{C},1} - \bar{Y}_{C,1}^{nat})$ and noting that the no-interference DiD uses $\bar{Y}_{C,1}^{nat}$ gives $\theta^{agg,*} = \theta^{agg} - \text{Bias}_L$. □

A1.5 Proof of Proposition 5 (Depletion Bias)

The argument is symmetric. Depletion affects the pre-treatment control mean: at $t = 0$, all agents (including future followers) are present, but at $t = 1$, followers have departed. The bias arises from the change in control-group composition between periods. The control-group first-difference under depletion differs from the undepleted case by $\text{Bias}_A = \frac{\phi_A N_{\mathcal{A}}}{N_{C,0}}(\bar{Y}_{\mathcal{A}\leftarrow\mathcal{C},0} - \bar{Y}^{\text{stay}}_{C,0})$, which captures the mechanical effect of removing selected agents from the control group at $t = 1$. □

A1.6 Proof of Proposition 6 (Decomposition under Interference)

Combine Propositions 2, 4, and 5. By Proposition 2, the no-interference aggregate DiD is $\theta^{agg} = \tau^{SATE} + \theta^{comp}$. Contamination subtracts Bias_L from the control-group post-treatment mean, and depletion subtracts Bias_A from the control-group first-difference. Since both operate on the control side and affect the DiD additively:

$$\theta^{agg,*} = \theta^{agg} - \text{Bias}_L - \text{Bias}_A = \tau^{SATE} + \theta^{comp} - \text{Bias}_L - \text{Bias}_A \tag{18}$$

□

A2 Lee-type bounds: derivation and two-sided extension

The Lee (2009) bounds rely on the monotonicity assumption: treatment weakly increases the probability of leaving. Under this assumption, the excess attrition $\hat{q}_L$ is attributable to treatment. Because the identity of these marginal leavers is unknown, bounds are obtained by considering worst-case selection: trimming the $\hat{q}_L$ fraction from each tail of the pre-treatment outcome distribution in turn yields samples consistent with the most and least favorable selection patterns.

One-sided case (exits only). To construct the lower bound on τ^{SATE} (making the effect smaller in magnitude), trim the $\hat{q}_L$ fraction of pre-treatment treated observations with the most extreme outcomes in the direction that would inflate the effect. For the upper bound, trim the opposite tail. Running the DiD on each trimmed sample yields bounds $\hat{\theta}^{LB} \leq \tau^{SATE} \leq \hat{\theta}^{UB}$.

Two-sided case (exits and entries). When both escapees and followers exist, define excess exit and entry rates $\hat{q}_L$ and $\hat{q}_A$. Trimming proceeds in two stages: $\hat{q}_L$ from the pre-treatment treated sample and $\hat{q}_A$ from the post-treatment treated sample. The four combinations yield:

$$\min\{\hat{\theta}^{LL}, \hat{\theta}^{LH}, \hat{\theta}^{HL}, \hat{\theta}^{HH}\} \leq \tau^{SATE} \leq \max\{\hat{\theta}^{LL}, \hat{\theta}^{LH}, \hat{\theta}^{HL}, \hat{\theta}^{HH}\} \tag{19}$$

The two-sided bounds assume that selection for exits and entries operates independently and that monotonicity holds for both margins. Standard errors are obtained via nonparametric bootstrap.

A3 Monte Carlo simulations

A3.1 Purpose and data-generating process

Each replication generates a two-period panel with $N = 1{,}200$ agents and $J = 60$ localities (20 agents per locality), split evenly between treated and control areas. The smaller sample size produces realistic estimation uncertainty for two-period DiD designs. Baseline heterogeneity enters through locality effects $\alpha_j \sim \mathcal{N}(0, 0.25^2)$ and individual types $u_i \sim \mathcal{N}(0, 1)$. Untreated outcomes follow:

$$Y_{i0}(0) = 1 + \alpha_{j(i)} + 0.6u_i + \varepsilon_{i0}, \qquad \varepsilon_{i0} \sim \mathcal{N}(0, 0.5^2), \tag{20}$$

$$Y_{i1}(0) = Y_{i0}(0) + 0.25 + 0.12\alpha_{j(i)} + 0.15u_i + \eta_i,\ \ \eta_i \sim \mathcal{N}(0, 0.25^2). \tag{21}$$

Treatment effects are heterogeneous within strata:

$$\tau_i \mid (i \in \mathcal{S}, u_i) \sim \mathcal{N}\big(\tau_{\text{base}} + 0.20u_i,\ 0.10^2\big), \tag{22}$$

$$\tau_i \mid (i \in \mathcal{L}, u_i) \sim \mathcal{N}\big(\tau_{\text{base}} + \delta + 0.20u_i,\ 0.10^2\big). \tag{23}$$

Migration is parameterized by $(\pi_L, \pi_A, \phi_L, \phi_A)$, with baseline values $(\pi_L, \pi_A) = (0.25, 0.20)$, $(\phi_L, \phi_A) = (0, 0)$, $\tau_{\text{base}} = -0.7$, $\delta = -0.6$. I run $R = 500$ replications per design cell and report Monte Carlo means with 5th–95th percentile bands.

A3.2 Results

Figure 2 in the main text shows the distribution of estimands under the baseline parameterization. The aggregate DiD is systematically shifted from the true ATT by the compositional effect, while the within-unit DiD tracks the true SATE.

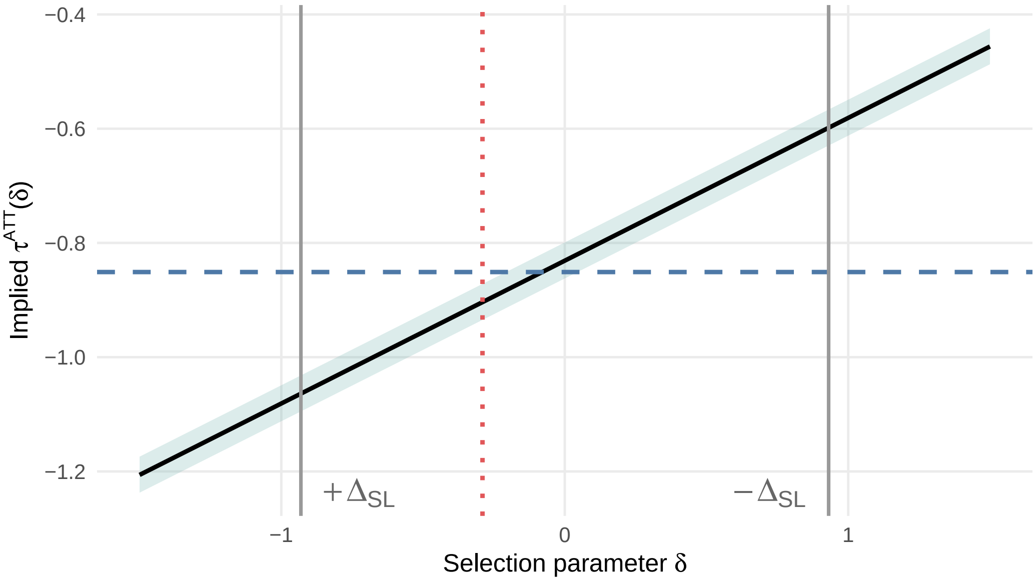
Figure A1 presents the ATT sensitivity curve. The implied ATT varies linearly with δ, with slope equal to the escapee share. Vertical markers at $\delta = \pm\hat{\Delta}_{SL}$ provide calibration anchors; the true ATT falls within the plausible range defined by the observable stayer–escapee gap.

Figure A2 validates the direct decomposition strategy. The estimated compositional effect $\hat{\theta}^{comp}$ closely tracks the true compositional effect across migration intensity levels, confirming that the decomposition is accurately estimated from sample statistics.

Figure A3 evaluates the Lee-type bounds. Under the monotone case ($\pi_A = 0$), bounds achieve correct coverage and widen with migration intensity. Under two-sided selection ($\pi_A > 0$), bounds are substantially wider.

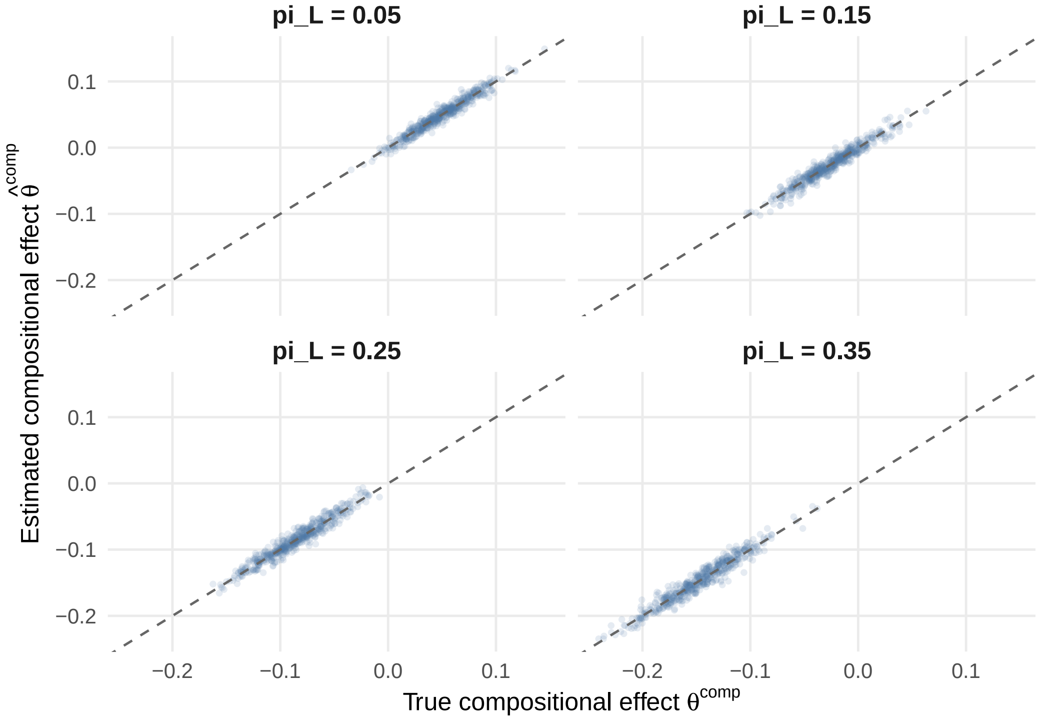
Figure A4 validates the observable stayer–escapee gap $\hat{\Delta}_{SL}$ as a calibration anchor for the sensitivity analysis. The left panel shows that $\hat{\Delta}_{SL}$ responds monotonically to the intensity of selection on levels: as the coefficient β governing how strongly the unobservable trait u predicts emigration increases, the pre-treatment gap between stayers and escapees widens. When $\beta = 0$ (random emigration), $\hat{\Delta}_{SL} \approx 0$; when $\beta = 1.2$ (strong selection), $|\hat{\Delta}_{SL}| \approx 1$. The right panel demonstrates how $\hat{\Delta}_{SL}$ anchors the sensitivity analysis in practice: vertical lines at $\delta = \pm\kappa\,|\hat{\Delta}_{SL}|$ for $\kappa \in \{0.5, 1, 2\}$ define nested plausible ranges for δ, in the spirit of Oster (2019). The key insight is that $\hat{\Delta}_{SL}$ provides an empirically grounded reference scale for δ without directly identifying it.

Figure A1: ATT sensitivity curve



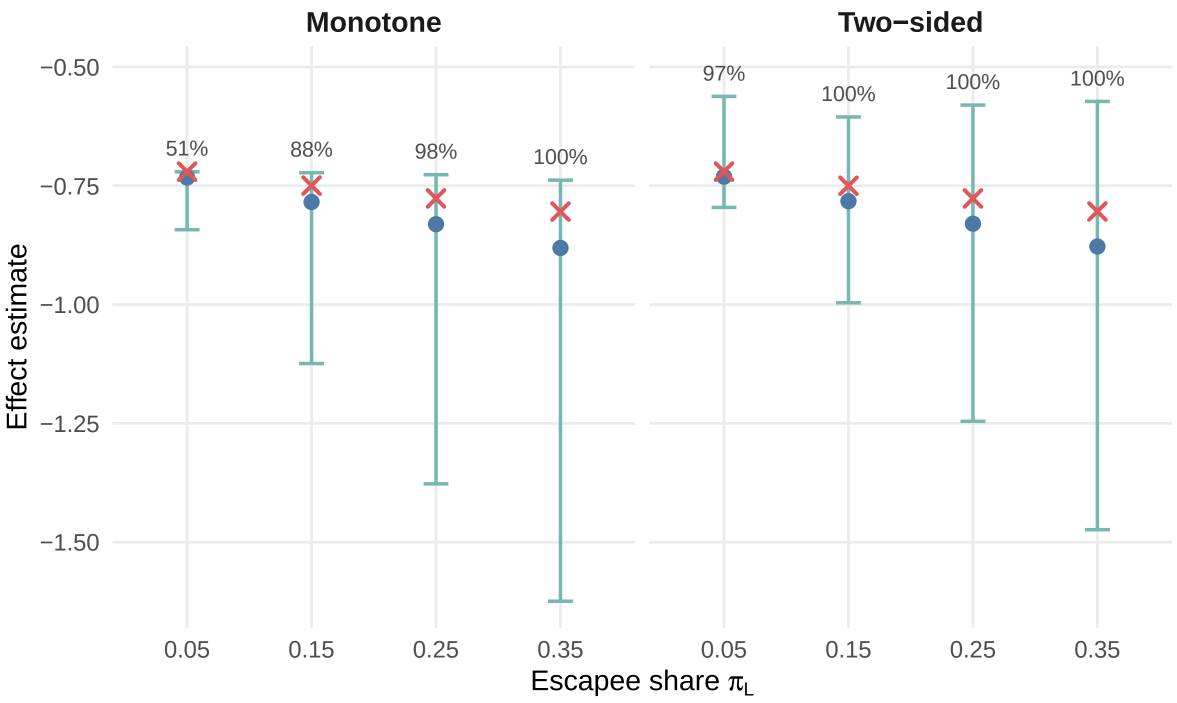
Note: The red line shows the implied $\tau^{ATT}(\delta)$ from Equation (11). The dashed blue line marks the mean true ATT; the dotted vertical line marks the mean true δ. The shaded region shows the 10th–90th MC percentile band. Vertical gray lines indicate the calibration anchors $\delta = \pm\hat{\Delta}_{SL}$.

Figure A2: Estimated vs. true compositional effect across migration intensity levels



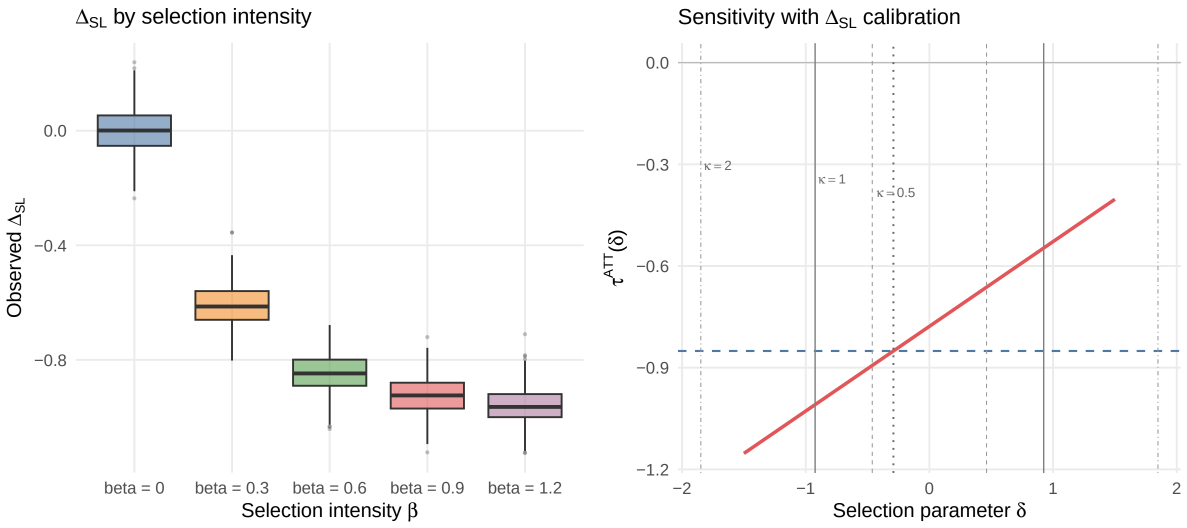
Note: The horizontal axis varies migration intensity ($\pi_L \in \{0.05, 0.15, 0.25, 0.35\}$). Each point is one Monte Carlo replication. The 45-degree line indicates perfect estimation.

Figure A3: Lee-type bounds across migration intensity levels



Note: Left: monotone selection ($\pi_A = 0$, no control turnover). Right: two-sided selection ($\pi_A = 0.15$). Blue dots show the Monte Carlo mean within-unit DiD; red crosses show the Monte Carlo mean true SATE; error bars show mean Lee bounds. Numbers indicate coverage rates. Under monotonicity, coverage improves with migration intensity as the trimming fraction becomes more precisely estimated.

Figure A4: Calibration anchor $\hat{\Delta}_{SL}$



Note: Left: $\hat{\Delta}_{SL}$ increases in magnitude with the selection intensity β (the coefficient on u in the emigration propensity score). When selection is random ($\beta = 0$), the stayer–escapee gap vanishes; stronger selection produces larger gaps. Right: the sensitivity curve $\tau^{ATT}(\delta)$ with calibration markers at $\delta = \pm\kappa\,|\hat{\Delta}_{SL}|$ for $\kappa \in \{0.5, 1, 2\}$ (using $\hat{\Delta}_{SL}$ from the $\beta = 0.9$ baseline). Dashed blue line: true ATT; dotted vertical line: true δ.

A4 Extension: Migration across Treatment and Control Areas

The main results require that treatment-induced migration does not cross the $\mathcal{T}$–$\mathcal{C}$ boundary (Assumption 1). The assumption might be violated whenever the treatment repels the escapees, or when it attracts followers. This appendix relaxes that assumption.

A4.1 Migration destinations

For each escapee $i \in \mathcal{L}$, define the destination $R_i \in \{\mathcal{C}, \mathcal{O}\}$, where $\mathcal{O}$ denotes outside the study area. The *control contamination rate* is $\phi_L \equiv \Pr(R_i = \mathcal{C} \mid i \in \mathcal{L})$. Symmetrically, for each follower $i \in \mathcal{A}$, define the origin $R_i \in \{\mathcal{C}, \mathcal{O}\}$. The *control depletion rate* is $\phi_A \equiv \Pr(R_i = \mathcal{C} \mid i \in \mathcal{A})$. Assumption 1 is equivalent to $\phi_L = \phi_A = 0$.

A4.2 Contamination by escapees

When $\phi_L > 0$, escapees enter $\mathcal{C}$ at $t = 1$. Let $N^*_{C,1} = N^{nat}_{C,1} + \phi_L N_{\mathcal{L}}$ denote the total observed control population at $t = 1$. The contaminated control post-treatment mean becomes a weighted average of native control agents and relocated escapees:

$$\bar{Y}^*_{C,1} = \frac{N^{nat}_{C,1}}{N^{nat}_{C,1} + \phi_L N_{\mathcal{L}}}\, \bar{Y}^{nat}_{C,1} + \frac{\phi_L N_{\mathcal{L}}}{N^{nat}_{C,1} + \phi_L N_{\mathcal{L}}}\, \bar{Y}_{\mathcal{L}\to\mathcal{C},1} \tag{24}$$

Proposition 4 (Contamination Bias). *When $\phi_L > 0$:*

$$\theta^{agg,*} = \theta^{agg} - \underbrace{\frac{\phi_L N_{\mathcal{L}}}{N^{nat}_{C,1} + \phi_L N_{\mathcal{L}}} \big(\bar{Y}_{\mathcal{L}\to\mathcal{C},1} - \bar{Y}^{nat}_{C,1}\big)}_{Bias_L} \tag{25}$$

The bias is positive (attenuating) when escapees have higher outcomes than control natives, and negative (amplifying) otherwise. It is negligible when the control group is large relative to the migrant flow ($N^{nat}_{C,1} \gg \phi_L N_{\mathcal{L}}$).

A4.3 Depletion by followers

When followers are drawn from $\mathcal{C}$ ($\phi_A > 0$), the control group loses agents selectively at $t = 0$. Let $N^{\text{stay}}_{C,0} = N_{C,0} - \phi_A N_{\mathcal{A}}$ denote the control agents who remain, and let $\bar{Y}_{\mathcal{A}\leftarrow\mathcal{C},0}$ denote the pre-treatment mean outcome of those who will become followers. The depleted control pre-treatment mean is:

$$\bar{Y}^*_{C,0} = \frac{N^{\text{stay}}_{C,0}}{N_{C,0}}\, \bar{Y}^{\text{stay}}_{C,0} + \frac{\phi_A N_{\mathcal{A}}}{N_{C,0}}\, \bar{Y}_{\mathcal{A}\leftarrow\mathcal{C},0} \tag{26}$$

Since the pre-treatment mean is computed before followers depart, $\bar{Y}^*_{C,0} = \bar{Y}_{C,0}$. The bias arises because these agents are absent from the control group at $t = 1$, distorting the control-group change.

Proposition 5 (Depletion Bias). *When $\phi_A > 0$:*

$$Bias_A = \frac{\phi_A N_{\mathcal{A}}}{N_{C,0}} \big(\bar{Y}_{\mathcal{A}\leftarrow\mathcal{C},0} - \bar{Y}^{stay}_{C,0}\big) \tag{27}$$

The bias is positive (attenuating) when followers drawn from the control group had higher pre-treatment outcomes than control stayers—removing high-Y agents depresses the control-group change, making the DiD look larger.

A4.4 Modified decomposition

Proposition 6 (Decomposition under Interference)**.** *When $\phi_L > 0$ or $\phi_A > 0$:*

$$\theta^{agg,*} = \tau^{SATE} + \theta^{comp} - Bias_L - Bias_A \tag{28}$$

The within-unit DiD is partially insulated from interference: stayers in treated localities are identified by individual fixed effects, and contamination affects $\hat{\theta}^{ind}$ only through the control-group trend. When the contamination share $\phi_L N_{\mathcal{L}}/N^*_{C,1}$ is small, the bias in the control trend is negligible and $\hat{\theta}^{ind}$ remains approximately unbiased for τ^{SATE}.

A4.5 Partial identification under sign restrictions

The biases Bias_L and Bias_A depend on unobservable counterfactual outcomes. However, sign restrictions yield informative bounds without point-estimating them (cf. Mealli & Viviens, 2025, Assumption 10).

Assumption 6 (Signed Contamination)**.** *$Bias_L \geq 0$, i.e., escapees who relocate to control localities have weakly higher outcomes than control natives: $\bar{Y}_{\mathcal{L}\to\mathcal{C},1} \geq \bar{Y}^{nat}_{C,1}$.*

This holds when escapees are positively selected relative to control units—for instance, when a minimum wage increase drives out low-productivity firms that nonetheless outperform typical control-area firms. Under Assumption 6:

$$\theta^{agg,*} \leq \tau^{SATE} + \theta^{comp} \tag{29}$$

That is, the contaminated aggregate DiD is a *lower bound* on the no-interference aggregate DiD. If additionally $\text{Bias}_A \geq 0$ (followers drawn from the upper tail of the control distribution), both biases attenuate and the bound tightens. The reverse sign restriction ($\text{Bias}_L \leq 0$) yields an upper bound instead.

Remark 1 (Magnitude restriction)**.** *A stronger assumption bounds the bias magnitude: $|Bias_L| \leq \bar{B}_L$. This yields a two-sided interval $[\theta^{agg,*} - \bar{B}_L,\ \theta^{agg,*} + \bar{B}_L] \ni \tau^{SATE} + \theta^{comp}$. The bound $\bar{B}_L$ can be calibrated from observable quantities. The migrant share $\phi_L N_{\mathcal{L}}/N^*_{C,1}$ is estimable when tracking data exists. The outcome gap can be bounded by the pre-treatment stayer–escapee gap $|\hat{\Delta}_{SL}|$ under the assumption that selection on levels is stable across periods.*

A4.6 Geographic buffers and donut controls

When researchers can track the complete population over time, it is possible to observe and exclude contaminating movements to the control areas. Otherwise, when migration is spatially concentrated near the treatment–control boundary, geographic distance provides an exclusion restriction for contamination (Butts, 2024).

Assumption 7 (Local Migration)**.** *There exists a distance threshold $\bar{d}$ such that no treatment-induced migrants relocate beyond it: for all control localities k with $\min_{j\in\mathcal{T}} d_{jk} > \bar{d}$, $\phi_L(k) = \phi_A(k) = 0$.*

Under Assumption 7, the "donut" control group $\mathcal{C}^{\bar{d}} = \{k \in \mathcal{C} : \min_{j\in\mathcal{T}} d_{jk} > \bar{d}\}$ is free of contamination. If additionally the parallel-trends assumptions (Assumptions 3–4) hold with $\mathcal{C}^{\bar{d}}$ as the control group, the

DiD using $\mathcal{C}^{\bar{d}}$ recovers the no-interference estimand θ^{agg} from Proposition 2. In practice, the researcher estimates the DiD across a sequence of buffer radii $\bar{d}_1 < \bar{d}_2 < \cdots$ and examines whether estimates stabilize beyond some threshold—a diagnostic for the contamination range.

When individual tracking data is available, Assumption 7 can be verified directly by checking whether treated-origin migrants appear in distant control localities. Even without tracking data, a sharp change in control-group growth rates near the boundary signals contamination. The difference $\hat{\theta}^{agg,\text{full}} - \hat{\theta}^{agg,\bar{d}}$ serves as a diagnostic proxy for the net contamination bias, under the assumption that the treated-side compositional effects and trend quality are comparable across the two control groups.

A4.7 Implications for estimation strategies

The three strategies from Section extend to the interference setting as follows.

Direct decomposition generalizes when migrant destinations are observed. In addition to classifying agents into stayers, escapees, and followers, the researcher tracks whether escapees relocate to control localities. The contamination bias Bias_L is then estimable from two quantities: the migrant share $\phi_L N_{\mathcal{L}}/N^*_{C,1}$ and the outcome gap $\bar{Y}_{\mathcal{L}\to\mathcal{C},1} - \bar{Y}^{nat}_{C,1}$. The depletion bias Bias_A requires observing which followers originated from control areas.

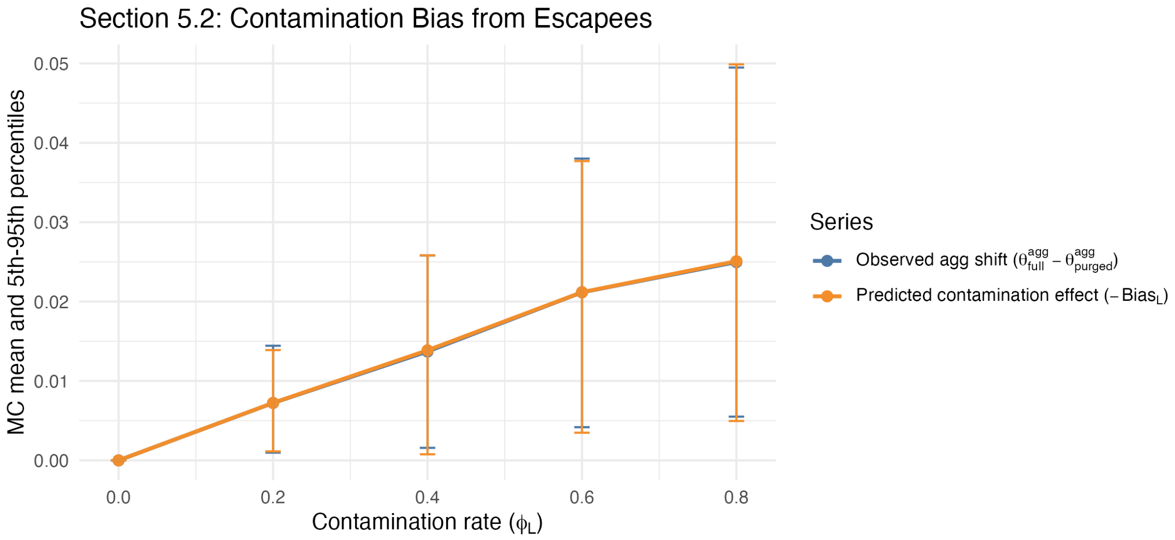
Lee-type bounds require a clean control group. When contamination is present, bounds should be computed using the donut control group $\mathcal{C}^{\bar{d}}$ from Assumption 7, which restores the no-interference setting at the cost of a potentially smaller and more distant comparison sample.

Sensitivity analysis for the ATT remains valid as a one-parameter exercise in δ when contamination is ruled out or bounded by sign restrictions. If the direction and magnitude of contamination are unknown, the analysis becomes a two-parameter problem in (δ, ϕ_L), with each parameter requiring separate calibration.

A4.8 Monte Carlo validation

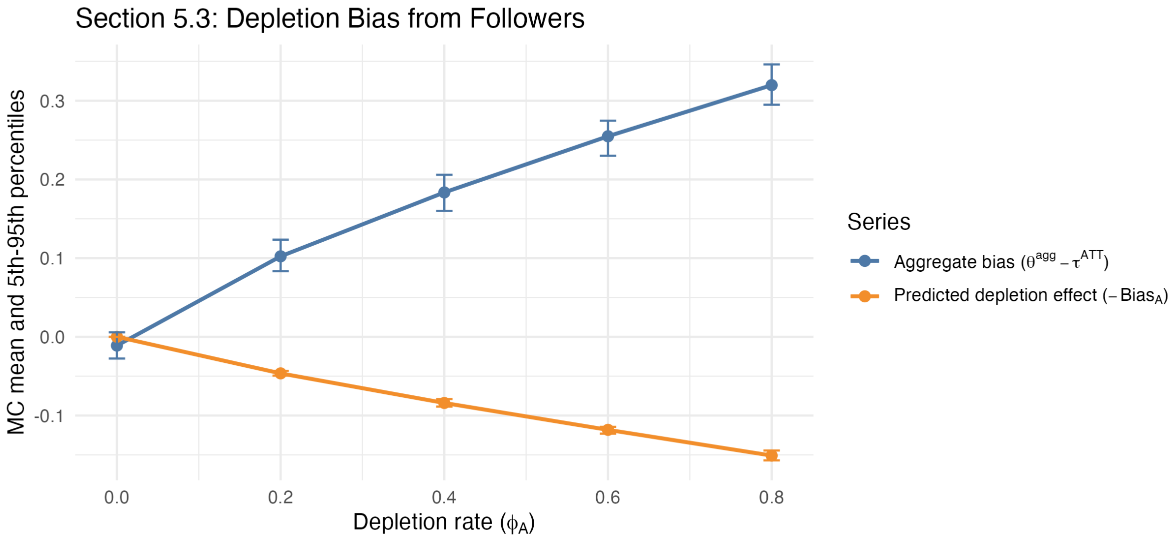
Figures A5–A7 validate the interference decomposition using the Monte Carlo design described in Appendix A3, extended to allow $\phi_L, \phi_A > 0$.

Figure A5: Contamination bias as ϕ_L varies



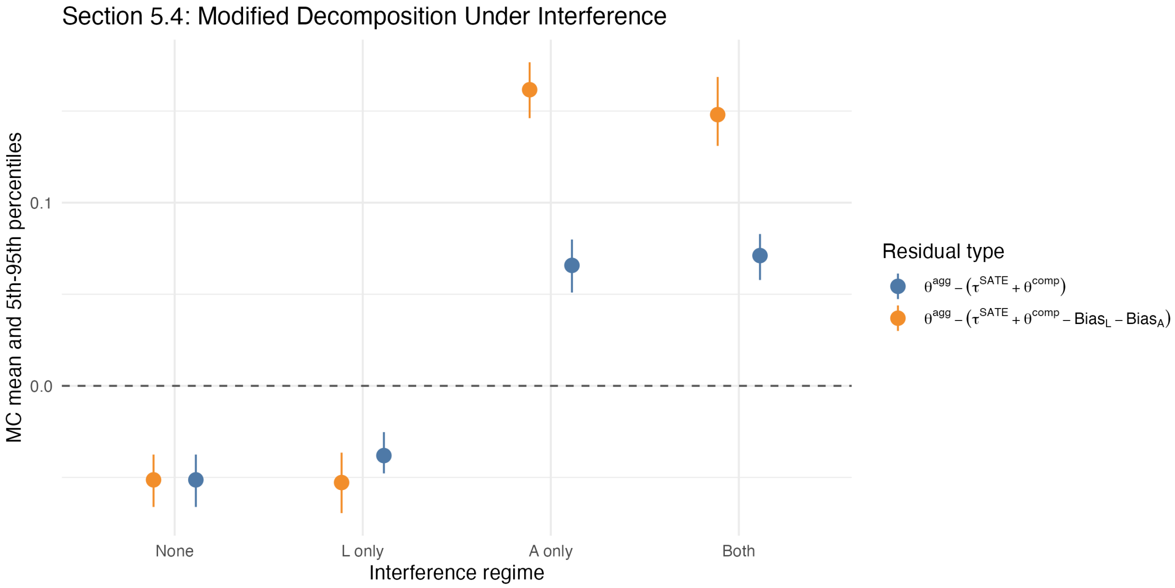
Note: Horizontal axis varies the contamination rate $\phi_L \in \{0, 0.2, 0.4, 0.6, 0.8\}$ with $\phi_A = 0$. Points show Monte Carlo means; error bars show 5th–95th percentile bands. The predicted contamination bias from Proposition 4 closely tracks the observed shift in the aggregate DiD.

Figure A6: Depletion bias from followers ϕ_A



Note: Horizontal axis varies the depletion rate $\phi_A \in \{0, 0.2, 0.4, 0.6, 0.8\}$ with $\phi_L = 0$. The predicted depletion bias from Proposition 5 closely tracks the observed shift.

Figure A7: Modified decomposition under joint interference



Note: Both $\phi_L > 0$ and $\phi_A > 0$. The no-interference residual $\theta^{agg} - \tau^{SATE}$ (left) is compared against the modified residual $\theta^{agg,*} - \tau^{SATE} + \text{Bias}_L + \text{Bias}_A$ (right), validating the additive structure of Proposition 6.

A5 Extension: Staggered Adoption with Monotone Exit Migration

This appendix provides an illustrative extension of the two-period setup to staggered treatment adoption with persistent migration risk. The goal is to characterize how mobility assumptions affect the bias between a naive DiD estimator and the cohort-weighted ATT in a multi-period rollout. For instance, imagine a mandatory vote policy that is gradually expanded over time and might lead part of the affected population to relocate to a control area. How does this migration affect the ATT estimates, and how does this depend on the number of times m agents are willing to move to escape the policy? To isolate the role of the move cap m, the treatment is deliberately simplified. As such, the setup assumes equal cohort weights, constant δ, and symmetric hazards.

A5.1 Setup: cohorts, exposure time, and estimands

Consider four periods $t \in \{0, 1, 2, 3\}$ with treatment expanded in three rounds. Let locality j have adoption cohort $G_j \in \{1, 2, 3, \infty\}$, where $G_j = \infty$ denotes never-treated localities. For treated cohorts, event time is $e = t - G_j$.

For each treated cohort-event cell (g, e), define:

$$ATT_{g,e} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid G_j = g,\, i \in \mathcal{S}_{g,e} \cup \mathcal{L}_{g,e}]\,, \tag{30}$$

$$SATE_{g,e} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid G_j = g,\, i \in \mathcal{S}_{g,e}]\,, \tag{31}$$

$$EATE_{g,e} = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid G_j = g,\, i \in \mathcal{L}_{g,e}]\,. \tag{32}$$

Let $\delta_{g,e} \equiv EATE_{g,e} - SATE_{g,e}$ denote selection on treatment effects.

Assume a per-exposure exit hazard q and move cap $m \in \{1, 2, 3\}$. After k exposure opportunities, the cumulative exit share is:

$$\omega_k(m) = 1 - (1 - q)^{\min(k,m)}. \tag{33}$$

Then the ATT–SATE relation becomes:

$$ATT_{g,e} = SATE_{g,e} + \omega_{e+1}(m)\, \delta_{g,e}. \tag{34}$$

Out of simplicity, here I do not consider the possibility that agents return to the treated areas at a later point.

A5.2 Migration process: contained one-move vs cascade

Scenario 1 (contained one-move monotone exit). Set $m = 1$. Each agent can move at most once in response to treatment and never returns. Then $\omega_k(1) = q$ for all $k \geq 1$, so the selection component of bias plateaus quickly in event time.

Scenario 2 (cascade monotone exit). Allow $m > 1$ under the same no-return restriction. Then $\omega_k(m)$ grows with exposure opportunities up to cap m, so earlier-treated cohorts accumulate larger selection bias by later calendar periods.

A5.3 Naive DiD with not-yet-treated controls (primary)

Define a period-level pooled naive DiD against control set C_t:

$$\theta_t^{naive,C} = \mathbb{E}[\Delta Y_{it} \mid D_{it} = 1] - \mathbb{E}[\Delta Y_{it} \mid i \in C_t], \tag{35}$$

where $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$. The target effect is the cohort-weighted ATT among currently treated localities:

$$ATT_t = \sum_{g \leq t} w_{g,t}\, ATT_{g,t-g}, \qquad \sum_{g \leq t} w_{g,t} = 1. \tag{36}$$

Using $C_t = NY_t$ (not-yet-treated) as the primary design:

$$\theta_t^{naive,NY} - ATT_t = B_t^{sel}(m) + B_t^{comp}(m) + B_t^{ctrl,NY}, \tag{37}$$

with:

$$B_t^{sel}(m) = -\sum_{g \leq t} w_{g,t}\, \omega_{t-g+1}(m)\, \delta_{g,t-g}, \tag{38}$$

$$B_t^{comp}(m) = \sum_{g \leq t} w_{g,t}\, \omega_{t-g+1}(m)\, \Delta_{SL,g}, \tag{39}$$

$$B_t^{ctrl,NY} : \text{ control-side composition term for not-yet-treated units.} \tag{40}$$

The composition-on-levels term $B_t^{comp}(m)$ follows from applying the two-period decomposition (Proposition 1) to each cohort-event cell: the escapee share $\omega_{t-g+1}(m)$ in cohort g at event time $t-g$ plays the role of $(1 - w_0)$, and $\Delta_{SL,g} = \bar{Y}_{\mathcal{S},g,0} - \bar{Y}_{\mathcal{L},g,0}$ is the pre-treatment stayer–escapee gap in cohort g. Under the simplification $\Delta_{SL,g} = \Delta_{SL}$ (constant across cohorts), this reduces to $B_t^{comp}(m) = \bar{\omega}_t(m) \cdot \Delta_{SL}$.

The control-side term $B_t^{ctrl,NY}$ captures compositional changes in the not-yet-treated control group. Under no contamination ($\phi_L = 0$) and no natural turnover in control localities, $B_t^{ctrl,NY} = 0$. When contamination is present, escapees from earlier-treated cohorts may relocate to not-yet-treated localities, introducing a time-varying control-side bias that grows as treatment expands geographically.

Proposition 7 (Monotonicity of Selection Bias in Move Cap)**.** *Under constant* $\operatorname{sign}(\delta_{g,e})$ *across* (g, e) *and a common per-exposure hazard* q*, the absolute selection bias* $|B_t^{sel}(m)|$ *is weakly increasing in the move cap* m *for all* t*. When* $m = 1$*,* B_t^{sel} *is constant in* t *(it plateaus after first exposure). When* $m > 1$*,* $|B_t^{sel}(m)|$ *grows with* t *as earlier-treated cohorts accumulate exits.*

The proof follows from $\omega_k(m) = 1 - (1-q)^{\min(k,m)}$, which is weakly increasing in both k and m. Since the weights $w_{g,t}$ are non-negative and $\delta_{g,t-g}$ has constant sign, $|B_t^{sel}(m)| = |\sum_g w_{g,t}\, \omega_{t-g+1}(m)\, \delta_{g,t-g}|$ is weakly increasing in m. Under $m = 1$, $\omega_k(1) = q$ for all $k \geq 1$, so $B_t^{sel}(1) = -q\, \bar{\delta}_t$ where $\bar{\delta}_t = \sum_g w_{g,t} \delta_{g,t-g}$, which does not grow with t when δ is constant. □

A5.4 Comparison with never-treated controls

Using never-treated controls gives:

$$\theta_t^{naive,N} - ATT_t = B_t^{sel}(m) + B_t^{comp}(m) + B_t^{ctrl,N}. \tag{41}$$

Therefore the estimand-level difference between control designs is:

$$\Delta B_t^{ctrl} \equiv B_t^{ctrl,NY} - B_t^{ctrl,N} = \left(\theta_t^{naive,NY} - ATT_t\right) - \left(\theta_t^{naive,N} - ATT_t\right). \tag{42}$$

When migration into control groups is negligible, ΔB_t^{ctrl} tends to be small and the two naive designs become closer.

A5.5 Relation to clean staggered estimators

Recent staggered-DiD estimators—Callaway and Sant'Anna (2021), Sun and Abraham (2021), Borusyak, Jaravel, and Spiess (2024)—address heterogeneous treatment effects and negative weighting by estimating cohort-specific effects $\widehat{ATT}(g,t)$ and aggregating them transparently. Under standard parallel trends and no anticipation, these estimators consistently estimate the cohort-specific $ATT_{g,e}$.

When treatment-induced migration is present, however, the decomposition in Proposition 2 applies to each cohort-event cell. A clean staggered estimator targeting cohort g at event time e identifies:

$$\widehat{ATT}(g,t) \xrightarrow{p} SATE_{g,e} + \omega_{e+1}(m) \cdot \Delta_{SL,g} + \text{control-side terms} \tag{43}$$

rather than $ATT_{g,e}$. Clean aggregation eliminates the heterogeneous-effects bias that contaminates naive two-way fixed effects, but does not address the compositional channel: each cell-specific estimate carries its own migration bias, which varies across cells through $\omega_{e+1}(m)$. The sensitivity analysis from Section applies cell by cell, with the escapee share and δ potentially varying by cohort.

A5.6 Simplified 4-period, 3-round illustration

Calibrate to the baseline migration scale in the paper: cumulative 3-opportunity exit share $\pi_L = 0.25$, implying

$$q = 1 - (1 - \pi_L)^{1/3} = 1 - 0.75^{1/3} \approx 0.091. \tag{44}$$

Use $\delta = -0.6$.

From $\omega_k(m) = 1 - (1-q)^{\min(k,m)}$:

$$\omega_1 = 0.091, \quad \omega_2 = \begin{cases} 0.091 & \text{if } m = 1, \\ 0.174 & \text{if } m \in \{2,3\}, \end{cases} \quad \omega_3 = \begin{cases} 0.091 & \text{if } m = 1, \\ 0.174 & \text{if } m = 2, \\ 0.250 & \text{if } m = 3. \end{cases}$$

At each period, active treated cohorts are:

$$t = 1 : \{1\}, \qquad t = 2 : \{1,2\}, \qquad t = 3 : \{1,2,3\}.$$

Define the average cumulative exit share among active treated cohorts:

$$\bar{\omega}_t(m) = \sum_{g \le t} w_{g,t}\, \omega_{t-g+1}(m). \tag{45}$$

Under equal cohort weights in this simplified illustration:

$$\bar{\omega}_1(m) = \omega_1, \quad \bar{\omega}_2(m) = \frac{\omega_2 + \omega_1}{2}, \quad \bar{\omega}_3(m) = \frac{\omega_3 + \omega_2 + \omega_1}{3}.$$

The primary bias components are:

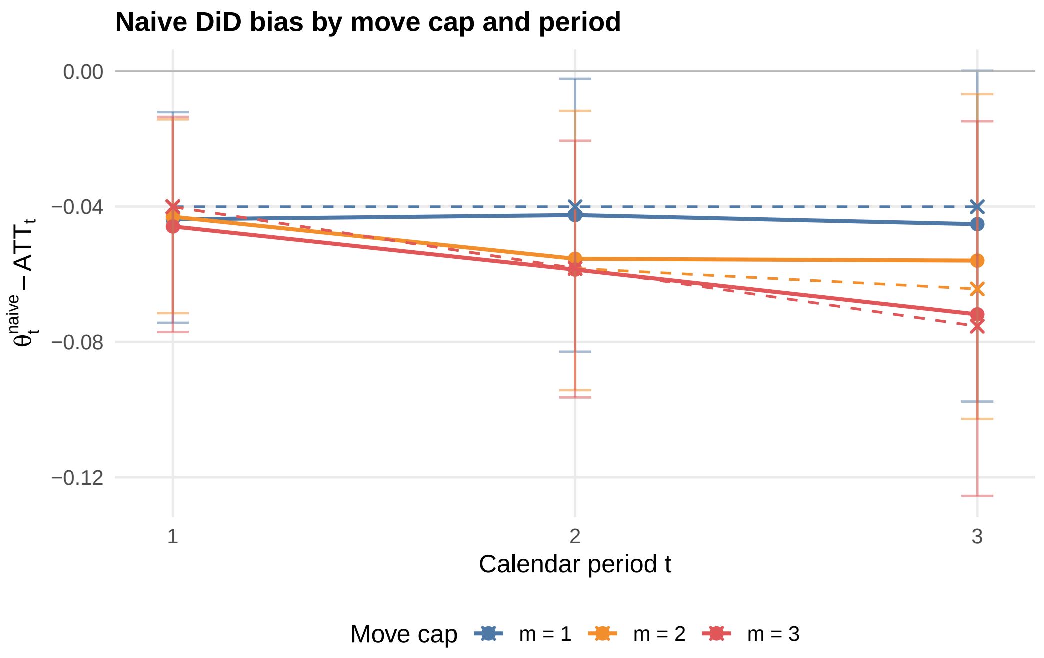
$$B_t^{sel}(m) = -\bar{\omega}_t(m)\,\delta, \qquad B_t^{comp}(m) = \bar{\omega}_t(m)\,\Delta_{SL}. \tag{46}$$

Both scale with the average cumulative exit share $\bar{\omega}_t(m)$ and therefore grow with the move cap m. Under $m = 1$, bias plateaus after the first exposure period; under $m > 1$, it accumulates with additional exposure.

Monte Carlo validation. To validate these predictions, we simulate a four-period panel ($t \in \{0, 1, 2, 3\}$) with $N = 1{,}200$ agents across $J = 60$ localities split equally among three treatment cohorts ($g = 1, 2, 3$) and a never-treated control group. The DGP matches the baseline two-period specification: $\delta = -0.6$, $q \approx 0.091$ (yielding $\pi_L = 0.25$ cumulative over three exposures), and selection on a latent factor u that enters both the escape propensity and baseline outcomes. Escapees leave the sample upon departure (no control contamination). We run $R = 200$ replications for each $m \in \{1, 2, 3\}$.

Figure A8 shows the total naive DiD bias $\theta_t^{naive} - ATT_t$ over calendar periods. Solid lines with 80% confidence bands are MC estimates; dashed lines with crosses show the analytical prediction $B_t^{sel} + B_t^{comp}$, using the MC-estimated Δ_{SL}. The analytical formulas closely track the simulation. Bias grows in magnitude with m: under $m = 1$ it stabilizes at approximately -0.045, while under $m = 3$ it reaches -0.072 by $t = 3$.

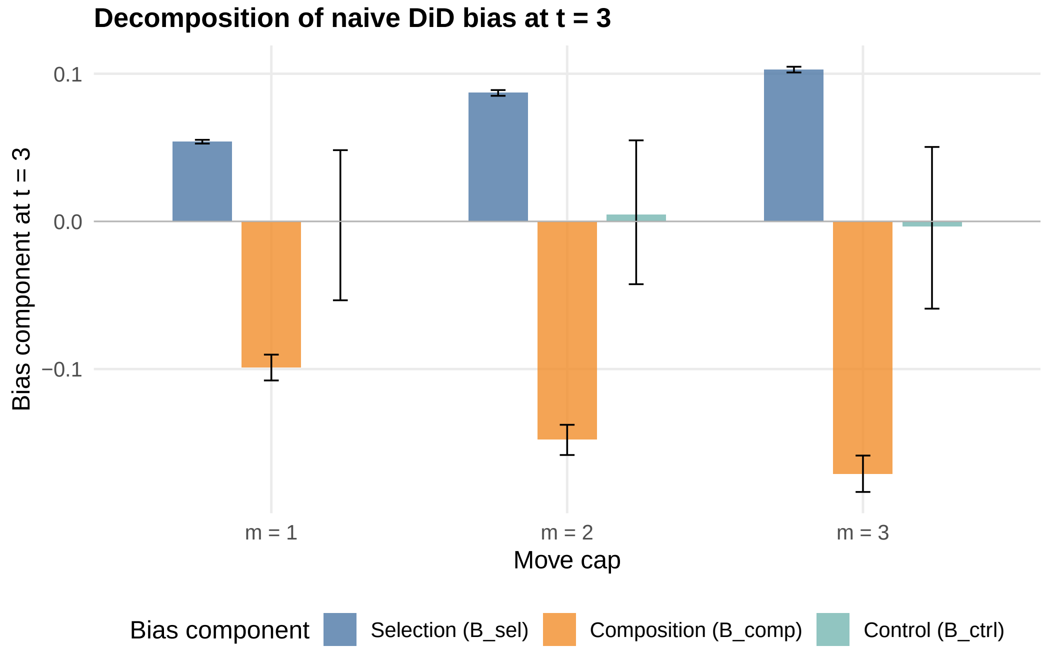
Figure A8: Naive DiD bias by move cap and period



Note: Solid lines: MC mean bias with 80% bands. Dashed lines with crosses: analytical prediction $B_t^{sel}(m) + B_t^{comp}(m)$.

Figure A9 decomposes the bias at $t = 3$ into three components. Selection on effects (B^{sel}, positive) reflects the gap between SATE and ATT: escapees have more negative treatment effects, so the stayer average overstates the population-level effect. Composition on levels (B^{comp}, negative) captures the shift in locality means when high-u escapees depart, making treated first differences appear more negative. In this DGP, B^{comp} dominates B^{sel}, producing a net negative bias. The control-side term (B^{ctrl}) is negligible, confirming that escapee departure does not contaminate the control group.

Figure A9: Decomposition of naive DiD bias by move cap



Note: Values at $t = 3$. B^{sel}: selection on effects (SATE ≠ ATT). B^{comp}: composition on levels (aggregate DiD ≠ SATE). B^{ctrl}: control-side contamination (approximately zero).

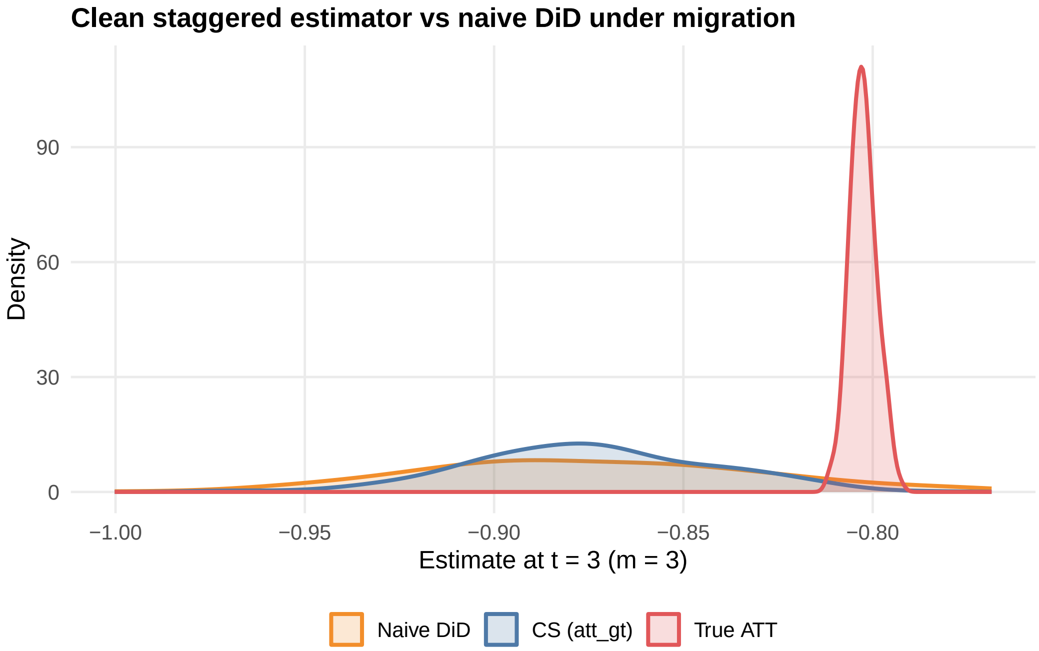
Figure A10 compares the naive aggregate DiD and the Callaway and Sant'Anna (2021) clean staggered estimator (CS) against the true ATT at $t = 3$ under $m = 3$. Escaped agents are dropped from the CS panel after departure, reflecting the realistic setting where outcomes of movers are unobserved. Both estimators exhibit comparable bias relative to the true ATT: CS is more efficient (tighter sampling distribution) but carries the same migration-induced bias, because the compositional and selection channels operate *within* each (g, e) cell and are not addressed by aggregation-robust estimation.

A5.7 Takeaways

1. If $q = 0$, then $\omega_k(m) = 0$ for all k and migration-induced bias vanishes.
2. If $\delta = 0$, then $ATT_{g,e} = SATE_{g,e}$ regardless of move cap.
3. With $m = 1$, both B^{sel} and B^{comp} are bounded and do not grow with longer exposure.
4. Under $m > 1$, bias accumulates with exposure time, creating heterogeneity across cohort-event cells that mimics treatment effect heterogeneity.
5. Clean staggered estimators (Callaway & Sant'Anna, 2021) address heterogeneous-effects bias across (g, e) cells but do not eliminate the within-cell compositional and selection channels.

This extension assumes monotone exit migration with no re-entry, constant δ and Δ_{SL} across cohorts, and no follower arrivals. Extending to two-way churn, cohort-varying selection parameters, and entry-side dynamics is left for future work.

Figure A10: True ATT vs. estimates from OLS and Callaway and Sant'Anna estimator



Note: Results at $t = 3$, with $m = 3$ and $R = 200$ replications. The clean staggered estimator is more efficient but does not remove selection or composition bias from migration.