sustainability

MDPI

Article

When Pilots Overlap: Contamination Bias in Difference-in-Differences Evaluations of China's City Pilot Policies for Sustainable Development

Chao Cheng [1] and Jinlin Zhang [2,*]

[1] School of Finance, Zhongnan University of Economics and Law, Wuhan 430073, China
[2] International Business School, Hainan University, Haikou 570228, China
* Correspondence: zjlifb@163.com

Abstract

China tests sustainability policies, from low-carbon and sponge cities to new-energy-vehicle promotion, in city pilot programmes, and their effects are usually estimated one programme at a time with staggered difference-in-differences (DID) designs. The programmes, however, overlap. Of 297 cities, 169 hosted two or more of ten national programmes by 2022, and 75–95% of each programme's treated city-years overlapped with another. Monte Carlo simulations calibrated to city PM2.5, with known effects for every programme, show that this overlap contaminates DID evaluations. When each co-located programme is as effective as the focal one, two-way fixed effects, event-study and heterogeneity-robust estimators overstate the focal effect by a median 78–139%. A programme with no effect is found significant in 11–22% of simulated evaluations when co-located effects are small (0.6%), and in 71–100% in a stress test that sets each to the median of three verifiable published estimates (−3.45%). Adding dummies for the other programmes, the usual robustness check, reverses the bias, and pretrend tests rarely detect the problem. Because the contamination is a weighted sum of the other programmes' effects, with weights fixed by the treatment pattern, these weights can be computed before outcome data are examined; we provide code and a checklist for evaluations of sustainability policies.

Keywords: sustainable development; sustainability policy evaluation; difference-in-differences; overlapping pilot policies; contamination bias; low-carbon city; PM2.5; Monte Carlo simulation; China

Check for updates

Academic Editor: Jianming Cai

Received: 10 October 2026
Revised: 1 November 2026
Accepted: 3 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

China's central ministries routinely test new policies in designated pilot cities before extending them nationwide [1,2]. Low-carbon cities, smart cities, e-commerce demonstration cities, sponge cities, new-energy-vehicle promotion cities and many similar programmes have each been evaluated in a large number of staggered difference-in-differences (DID) studies (e.g., [3]). Many of these programmes, from low-carbon and sponge cities to new-energy-vehicle promotion, are instruments of China's environmental and low-carbon policy, and their evaluations are meant to inform which measures are extended nationwide. These studies compare cities selected in successive batches with cities not yet or never selected. The appeal is clear: the batches look like a quasi-natural experiment, and city-level outcomes are easy to obtain.

The difficulty is that the experiments are not run one at a time. Programmes are assigned by different ministries, but they tend to land in the same places: provincial capitals,

sub-provincial cities and the coastal east. In the database we assemble for ten national programmes, 169 of 297 cities hosted two or more programmes by 2022, and one city hosted all ten. Between 75% and 95% of each programme's treated city-years were also treated by at least one other programme. A DID estimate for one pilot therefore compares cities that differ in several pilots at once. When one programme is the focus of an evaluation, we call the other programmes co-located programmes, a treated city-year that also hosts one of them overlapped, and the part of the estimate that reflects their effects contamination.

The problem is not peculiar to China. Evaluators routinely estimate the effect of one programme in jurisdictions exposed to several: US states adopt marijuana policies as bundles that keep changing after adoption [4], state policies co-occur in ways that complicate the estimation of any one of them [5], and housing price controls are introduced, tightened, relaxed and abolished in sequence [6]. Chinese city pilots are an unusually clear case: the programmes are numerous, their designation dates are public, and the overlap among them can be measured exactly.

Applied researchers know this. A robustness check titled "excluding the interference of other policies" has become a standard part of the genre. In ten recent open access studies whose full texts we examined, seven added dummy variables for other pilots or policy events to the regression and three deleted cities that hosted other pilots. To our knowledge, whether these checks work has not been examined systematically.

The recent econometric literature on DID does not settle the question, because it addresses a different problem. It shows that two-way fixed effects (TWFE) regressions can be badly biased when the effect of a single staggered treatment varies across adoption cohorts and over time, and it offers estimators that avoid this bias [7–11]. These estimators clean up the comparison for one treatment. They do not remove the influence of other treatments that differ between the groups being compared. Recent theory shows that regressions with several treatments can be contaminated by the effects of the other treatments [12,13]. How large this contamination is in a real overlap structure, and what it does to the practices applied researchers actually use, has, to our knowledge, not been quantified.

This paper measures it. We combine the observed overlap of ten Chinese city pilot programmes with Monte Carlo simulations in which the true effect of every programme is known. Outcomes are calibrated to city-level PM2.5 concentrations, a common outcome in this literature, and a half-synthetic design keeps the satellite-derived PM2.5 paths. For each of six focal pilots, we evaluate fourteen estimators and robustness practices across 65 designs: TWFE, an event study, four heterogeneity-robust estimators, three ways of adding controls for other pilots, four ways of deleting observations, and an interaction model. We consider two effect sizes: a small one, and one equal to the median of three verifiable published estimates, a 3.45% reduction in PM2.5. Giving this benchmark to every co-located programme makes the second a stress test.

Four findings stand out. First, every estimator that ignores the other pilots is contaminated, and the heterogeneity-robust estimators are no exception. When each co-located programme has an effect as large as the focal one, the median estimate overstates the focal effect by 78–139%, depending on the estimator. Whether this bias produces false positives depends on how large the co-located effects are relative to the noise. With small effects, the same estimators reject a true null of no effect in 11–22% of simulated samples; in the stress test with literature-sized effects, they reject it in 71–100%. In the reference design, the bias that heterogeneity-robust estimators were built to remove is small, below 5% of the effect for TWFE, and it stays below 11% in every design without interactions between programmes. The bias they leave in place is large.

Second, the standard robustness check does not fix the problem. Adding dummy variables for the other pilots removes their contamination only when each of those pilots

has the same effect in every treated city. When their effects vary across cities in a way that lines up with the focal pilot's rollout, the bias changes sign instead of disappearing. Estimates then understate the focal effect by 48–66%, and with literature-sized effects a true null is rejected in 50–69% of samples. Which failure dominates, overstatement without the dummies or understatement with them, depends on a feature of the other pilots' effects that researchers cannot observe.

Third, pretrend tests offer little protection. In contaminated designs that pass a pretrend test, the false-positive rate with literature-sized effects is still 82–85%, against 5% in clean designs. Fourth, designs that keep only comparisons in which neither treated nor comparison cities host other pilots remove most of the contamination. They leave between 2 and 16 treated cities, however, and they change the estimand. An interaction model that represents overlap explicitly keeps the false-positive rate near 9%, at the cost of a downward bias of 26–30%. No method in our menu recovers the full-sample average effect free of contamination.

Behind these results is a simple fact: all the estimators we study are linear in the outcome. The contamination they carry is therefore a weighted sum of the other pilots' effects, with weights that depend only on the treatment matrix and the estimator. We compute these weights for every estimator and focal pilot and show that they reproduce the simulated bias almost exactly for TWFE, the event study and the heterogeneity-robust estimators. Because the weights can be computed before any outcome data are examined, they give applied researchers a practical diagnostic: run the estimator on synthetic outcomes equal to each co-located programme's assumed effect and see how far the estimate moves.

An illustration with satellite-derived PM2.5 shows what is at stake. It does not estimate the programmes' effects; it shows how much estimates vary across specifications and how large contamination could be under stated assumptions. Standard estimators find significant PM2.5 reductions of about 2–3% for the low-carbon, sponge-city and cross-border e-commerce pilots. If each co-located programme reduced PM2.5 by the benchmark 3.45%, contamination alone would shift the TWFE estimate for low-carbon cities by 0.054–0.061 log points (5.3–5.9%), about three times the estimate itself. Estimators that limit contamination give imprecise and sometimes opposite-signed estimates.

The paper contributes to three bodies of literature. The first is the econometrics of DID and regression with several treatments [12–14]. We show that, in simulations built on a real overlap structure, contamination from co-located treatments rather than heterogeneity of the focal effect is the first-order problem, and that the common fixes fail in ways the theory predicts. The second is work on co-occurring policies. Griffin et al. [5] use simulations to study two co-occurring state policies in the United States, and Yang and Tol [15] address contamination bias among overlapping environmental markets in China with firm-level data. We use the full overlap matrix of ten programmes, evaluate the menu of practices researchers actually use, and calibrate the outcome to the setting in which these practices are most common. The third is the large applied literature on Chinese pilot policies, including work on synergies between urban pilots [16] and on pairs of pilots evaluated jointly [17]. Our results imply that estimates which ignore co-located programmes, or merely add dummies for them, can be dominated by contamination.

For evaluators, the main practical contribution is a design-stage diagnostic and a checklist for reporting it. Like power calculations for DID [18] and simulation-based checks of interrupted time series designs [19], the diagnostic can be run before any outcome data are examined, which also makes it easy to include in a pre-analysis plan [20].

Section 2 describes the pilot programmes, their overlap and current practice. Section 3 shows why overlap biases DID estimators and defines contamination weights. Section 4 sets out the simulation design. Section 5 reports the results, Section 6 the empirical illustration, and Section 7 the implications for evaluation practice. Section 8 concludes.

2. Overlapping Pilots and How Applied Studies Handle Them

2.1. Ten National Pilot Programmes

We code ten national programmes that designate prefecture-level cities in batches and that are widely evaluated with DID: low-carbon city pilots, innovative city pilots of the National Development and Reform Commission, cross-border e-commerce comprehensive pilot zones, e-commerce demonstration cities, new-energy-vehicle promotion cities, sponge city pilots, social credit system demonstration cities, energy-saving and emission-reduction fiscal demonstration cities, national civilised cities, and Broadband China demonstration cities. Appendix Table A1 lists the programmes with their treated cities and overlap statistics.

The panel covers 297 cities at or above the prefecture level, including the four municipalities, from 2000 to 2022, or 6831 city-years. Each of the 760 first designations of a city into a programme is taken from the official list and recorded with its source, year and mapping rule. A city is treated from the year of its official designation if the city as a whole or any of its urban districts is designated; designations of counties and county-level cities are not extended to the prefecture. Joint designations of city clusters are expanded to their member cities; designations of whole provinces, such as the low-carbon pilot provinces, are not extended to their cities, and new areas and autonomous prefectures lie outside the panel. Where lists overlap or were revised, a city enters a programme at its first designation and remains treated: later reviews, renewals and withdrawals are not coded, and a designation is dated by the year of the document that made it, not the year it was posted online. Related but separate programmes, such as the innovative-city lists of the Ministry of Science and Technology, the 2009–2010 new-energy-vehicle demonstration and the 2021 sponge-city demonstration, are not merged into the ten programmes. Later national batches that we could not verify are not coded as treatment in the main analysis. Sixty-four designations are flagged as awaiting final confirmation: 49 sponge-city and energy-saving designations taken from official publicity lists whose final funding lists we could not verify, and 15 low-carbon designations of 2017 approved subject to revision of the cities' plans. Appendix Table A3 summarises the batches, issuing authorities, mapping rules and uncertain or excluded cases of each programme, and Supplementary Table S1 gives the official document, web address and retrieval date of each batch. Section 5.6 reports the overlap under three alternative codings, including one that drops the flagged designations.

2.2. How Much the Programmes Overlap

Overlap is the rule rather than the exception (Figure 1). By 2022, 238 of the 297 cities hosted at least one programme, 169 hosted two or more, and one city hosted all ten. The average number of programmes per city rose from 0.16 in 2010 to 1.27 in 2015 and 2.56 in 2022. The share of cities with three or more programmes rose from below 1% to 42% over the same period.

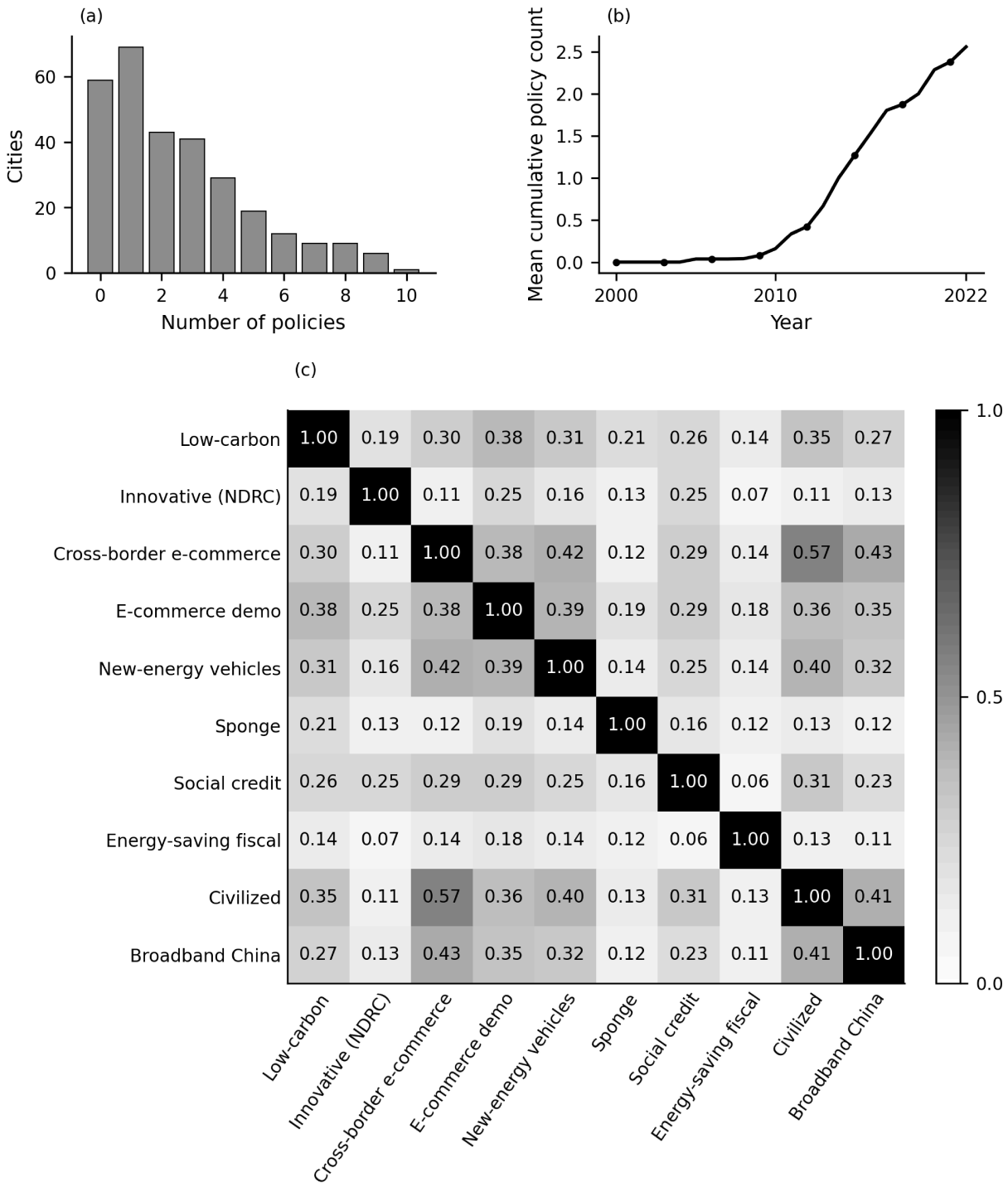


Figure 1. Overlap of ten national pilot programmes, 297 cities, 2000–2022. Panel (**a**): number of programmes per city in 2022. Panel (**b**): mean number of programmes per city by year. Panel (**c**): Jaccard index of the sets of designated cities for each pair of programmes, that is, the number of cities in both sets divided by the number in either.

The programmes concentrate in the same kind of city. Sub-provincial cities hosted 7.5 programmes on average and the four municipalities 8.5, against 2.0 for ordinary prefecture-level cities. Eastern cities hosted 3.9 programmes on average, against 1.6 to 2.3 in the other three regions. All 17 innovative pilot cities were also cross-border e-commerce zones, e-commerce demonstration cities and civilised cities, and 56 of the 70 low-carbon pilot cities were civilised cities.

https://doi.org/10.3390/su182111300

What matters for a DID evaluation is overlap within the treated observations. For each programme, between 75% (Broadband China) and 95% (social credit) of its treated city-years were also treated by at least one other programme. A treated city-year carried 2.5 to 4.6 other programmes on average. The overlap is not an artefact of our coding: dropping the flagged designations leaves 160 cities with two or more programmes, and treating unverified later national batches as treated raises the figure to all 297.

2.3. How Applied Studies Deal with Overlap

Robustness sections titled "excluding the interference of other policies" are a regular feature of DID studies of Chinese pilots. To see what they do, we examined ten recent open-access studies whose full texts could be verified. This is a purposive sample, not a systematic review. Seven of the ten add dummy variables for between two and four other programmes or policy events to the TWFE regression [21–27]. The other three delete cities that host other programmes [28–30]. One of them states that it removes these cities from both the treated and the comparison group; the other two do not say clearly which group is cleaned.

These practices correspond closely to estimators in our simulations, although two of the deletion rules are not described precisely enough to match a single design. Adding dummies corresponds to TWFE with controls for the other programmes, and the deletion rules correspond to our sample-cleaning designs. Section 3 shows why each is or is not expected to remove contamination, and Section 5 measures how well each performs.

3. Why Overlap Biases DID Estimators

3.1. Setup

Consider cities $i = 1, \ldots, N$ observed in years $t = 1, \ldots, T$, and K programmes with absorbing treatment indicators D_{kit}. The outcome is

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k=1}^{K} D_{kit}\,\tau_{kit} + \varepsilon_{it}, \tag{1}$$

where τ_{kit} is the effect of programme k in city i and year t. The effect may vary with time since adoption, with the adoption cohort and across cities. The untreated shocks ε_{it} satisfy parallel trends. For a focal programme f, the target is the average effect on the treated city-years, ATT_f, the mean of τ_{fit} over all city-years with $D_{fit} = 1$, holding the other programmes at their actual status. These other programmes are the co-located programmes of f.

3.2. Linear Estimators and Contamination Weights

Every estimator we study is linear in the outcome. TWFE with or without controls, aggregated event-study coefficients, the estimators of Callaway and Sant'Anna [9], Sun and Abraham [10], Borusyak et al. [11] and de Chaisemartin and D'Haultfœuille [31], stacked regressions, sample-deletion designs and the interaction model can all be written as $\hat{\beta}_f = \Sigma_{it}\, a_{it}\, Y_{it}$. The weights a_{it} are fixed by the treatment matrix and by the estimator's sample rules, and they sum to zero within each city and each year, so city and year effects drop out. Substituting (1) gives

$$\mathrm{E}\left[\hat{\beta}_f\right] - \mathrm{ATT}_f = \underbrace{\sum_{i,t} a_{it} D_{fit}\,\tau_{fit} - \mathrm{ATT}_f}_{\text{own-effect weighting}} + \underbrace{\sum_{k \neq f} \sum_{i,t} a_{it} D_{kit}\,\tau_{kit}}_{\text{contamination}} \tag{2}$$

The first term is the bias that the recent DID literature addresses. TWFE can weight some focal effects negatively when effects vary across cohorts and over time [7,8], and heterogeneity-robust estimators choose a_{it} so that this term vanishes. The second term

is the contamination from the other programmes. Nothing in the construction of the heterogeneity-robust estimators constrains it.

Write the effect of programme k as $\tau_{kit} = \mu_k \, h_{kit}$, where μ_k is its average effect over its own treated city-years and the path h_{kit} has mean one over the same city-years. The contamination weight of programme k for path h is

$$w_k(h) = \sum_{i,t} a_{it} \, D_{kit} \, h_{kit}, \qquad \text{contamination} = \sum_{k \neq f} \mu_k \, w_k(h_k). \tag{3}$$

A weight $w_k = 0.5$ means that if programme k reduced the outcome by 1% on average, the estimate for programme f would move by 0.5 percentage points. The sum $\Sigma_k \, w_k$ gives the response of the estimate when all co-located programmes have effects of the same sign and size.

3.3. What Common Practices Do to the Weights

Ignoring the other programmes. TWFE, event studies and heterogeneity-robust estimators place no restriction on $\Sigma_{it} \, a_{it} \, D_{kit}$. The weights are non-zero whenever programme k is more common, or arrives at different times, among the observations the estimator treats as treated than among those it uses as comparisons. Heterogeneity-robust estimators compare focal-treated cities with cities not yet or never treated by the focal programme. Those comparison cities may host other programmes too, but less often, so the contamination remains.

Adding dummies for the other programmes. By the Frisch–Waugh–Lovell theorem, controlling for D_{kit} forces $\Sigma_{it} \, a_{it} \, D_{kit} = 0$, so $w_k = 0$ when programme k has the same effect in every treated city-year ($h_{kit} = 1$). For any other path, $w_k(h) = \Sigma_{it} \, a_{it} \, D_{kit} \, (h_{kit} - 1)$: the zero-sum weights multiply the deviations of programme k's effect from its average. The result can take either sign and is the contamination bias analysed by Goldsmith-Pinkham et al. [13] and de Chaisemartin and D'Haultfœuille [12]. Dynamic dummies for programme k remove contamination from an effect path common to all of its cohorts, and cohort-by-treatment dummies remove cohort-specific effects. None of the dummy sets used in practice, whether static, dynamic or cohort-specific, removes city-specific deviations that are correlated with the focal programme's rollout.

Deleting observations. If every observation with a non-zero weight has $D_{kit} = 0$, then $w_k = 0$ for any path. This requires that both the treated and the comparison observations be free of other programmes. Deleting overlapped cities from only one group does not achieve it and can enlarge the weights. The price of full cleaning is a smaller sample and a different estimand: the effect in cities without other programmes, or in the years before other programmes arrived.

Modelling the overlap. An interaction model adds terms $D_{fit} \times D_{kit}$ to the TWFE regression and reports the focal effect averaged over the observed combinations of programmes. It removes contamination when effects are constant within each combination, but it is misspecified when they vary over time or across cities.

3.4. From Bias to False Positives

Let δ denote the bias of $\hat{\beta}_f$ divided by its standard error. With approximately normal estimates and correctly estimated standard errors, a nominal 5% test of $\beta_f = 0$ rejects a true null with probability

$$\Pr(\text{reject}) \approx \Phi(-1.96 - \delta) + \Phi(-1.96 + \delta). \tag{4}$$

By (3), the bias grows in proportion to the effects of the co-located programmes, while the standard error is governed by the noise in the outcome. The relative bias, bias divided by ATT_f, depends on the ratio of co-located to focal effects. The false-positive rate depends

on co-located effects relative to the noise. A design that produces a modest size distortion when co-located effects are small can therefore produce near-certain false positives when they are as large as published estimates.

3.5. A Diagnostic Available Before Estimation

Because the weights a_{it} depend only on the treatment matrix, $w_k(h)$ can be computed before any outcome is examined. One runs the chosen estimator on a synthetic outcome equal to $D_{kit}\, h_{kit}$, with no noise, for each co-located programme k and each path h of interest. The estimate is $w_k(h)$. Multiplying by a plausible range for μ_k gives the bias that co-located programmes could induce. The weights thus measure how sensitive an estimate is to the co-located effects; they do not identify the bias in a given application, which also depends on effects that are not observed. Because the weights are signed, co-located effects of the opposite sign to the focal effect reverse the contamination, and effects of mixed signs partly offset. Section 5.5 shows that this calculation reproduces the simulated bias, and Section 6 applies it to observed data.

4. Simulation Design

4.1. Treatment Matrix and Focal Programmes

The main grid and the literature-scale designs use the observed treatment matrix of Section 2, held fixed across replications: 297 cities, 23 years (2000–2022) and 10 programmes. The half-synthetic design (Section 4.5) and the empirical illustration (Section 6) require observed PM2.5 and therefore use the 289 cities and 22 years (2000–2021) for which it is available, with the treatment matrix restricted to these cities and years. Overlap is almost as pervasive in this shorter panel: 148 cities host two or more programmes by 2021, and 74–95% of each programme's treated city-years overlap with another programme. Most of the difference from the full panel comes from the final year, in which 53 cities became cross-border e-commerce zones, rather than from the eight cities without PM2.5 data. Six programmes serve in turn as the focal programme: low-carbon cities, e-commerce demonstration cities, new-energy-vehicle cities, sponge cities, civilised cities and cross-border e-commerce zones. They cover a wide range of overlap, with 79–93% of treated city-years overlapped, and of size, from 27 to 152 treated cities. When one programme is focal, the other nine are co-located programmes with effects of their own.

4.2. Outcome Model

Simulated outcomes follow Equation (1), extended with features of city air-quality data. Log PM2.5 is the sum of city effects, year effects, city-specific linear trends, province-by-year shocks, the effects of all ten programmes and an AR(1) idiosyncratic error. The components are calibrated to the area-weighted annual log PM2.5 of 289 cities in the ChinaHighPM2.5 product for 2000–2021 [32,33]. The other eight cities (Hefei, Wuhu, Ma'anshan, Suzhou, Yingkou, Neijiang, Danzhou and Sansha) did not meet the pre-registered rules for matching city boundaries to the PM2.5 grid. The calibration fixes only the variance components and the error process; the simulated panels cover all 297 cities. The idiosyncratic component has standard deviation $\sigma = 0.0284$ log points (about 2.9%) and first-order autocorrelation 0.356. Separate designs add spatial spillovers from treated neighbours within 200 km, anticipation in the year before designation, and selection of cities on their underlying trends.

4.3. Programme Effects

The effect of each programme is $\tau_{kit} = \mu_k\, h_{kit}$, as in Section 3. The path h_{kit} is homogeneous, cohort-specific or dynamic. Dynamic paths rise with years since designation at a step $s \in \{0, 0.05, 0.15, 0.30\}$ for five years and then stay constant. On top of these, city-specific

deviations in the co-located programmes' effects are correlated with the focal programme's adoption cohort, with correlation $c \in \{0, 0.2, 0.4, 0.6, 0.8\}$. A positive c means that cities designated early for the focal programme also experience systematically different effects of the other programmes. Every path is rescaled so that its mean over the programme's own treated city-years is one, which makes μ_k the programme's average effect. The type of heterogeneity applies to all programmes, the focal one included: in homogeneous designs, every effect is constant; in cohort designs, every effect is cohort-specific; and in dynamic designs, the focal effect rises with step 0.15 while the co-located step s varies.

We use two effect scales. On the main scale, the focal effect is $\mu_f = 0.2\sigma = 0.0057$ log points (about 0.6%), with variants at 0.1σ and 0.3σ. Each co-located programme has $\mu_k = r \times 0.2\sigma$ with $r \in \{0, 0.5, 1, 2\}$. On the literature scale, the focal effect is $\mu_f = -0.0351$ log points (-3.45%), the median of the three log-PM2.5 DID coefficients for Chinese city pilots whose published full texts we could verify [28,30,34] (Appendix Table A2). Each co-located programme has $\mu_k = r \times (-0.0351)$ with $r \in \{0.5, 1, 2\}$. The literature scale is a stress test: it assumes every co-located programme is as effective as this benchmark. Because all estimators are linear, the sign convention does not matter: reversing all effects reverses the bias and leaves relative bias and rejection rates unchanged.

4.4. Designs and Replications

The main grid has 65 designs per focal programme. Its core crosses the five values of c with the four dynamic steps, each with a non-zero and a zero focal effect (40 designs). The remaining 25 designs vary the type of heterogeneity, the ratio r, the focal scale, interactions between programmes (the focal effect changes by $\pm 0.1\sigma$ per additional co-located programme), autocorrelation, selection, spillovers and anticipation, and include a benchmark with no co-located effects. With six focal programmes, this gives 390 cells, each simulated 500 times when the focal effect is non-zero and 1000 times when it is zero, a total of 261,000 panels.

On the literature scale, low-carbon and e-commerce cities are focal in 36 cells that cross $r \in \{0.5, 1, 2\}$, $c \in \{0, 0.4, 0.8\}$ and a zero or non-zero focal effect (27,000 panels). The other four focal programmes are simulated in a reference design (6000 panels). A benchmark with the same focal effects but no co-located effects provides a clean comparison for pretrend tests.

Throughout, the reference design has dynamic effects with step 0.15, moderate alignment ($c = 0.4$), co-located effects equal in size to the focal effect ($r = 1$), and no interactions, spillovers, anticipation or selection. The appendix reports every design. The calibration, the design grid, the estimator list, the random seeds and the analysis rules were fixed in a pre-analysis plan that was frozen by commit hash before the formal runs and is included in the replication package, in the spirit of Dee's case [20] for preregistering quasi-experimental evaluations. Analyses added after the formal runs are exploratory, and the replication package lists them as such. The pre-specified results are those of the 65-design grid and the half-synthetic design on the main scale (the main-scale results in Section 5 and in Appendix Tables A4, A5 and A8). The exploratory results are all those on the literature scale, including panel (b) of the two figures in Section 5.4; the contamination weights and the sensitivity calculations based on them (Section 5.5 and Appendix Tables A6 and A9); the additional dCDH runs (Table A7); the empirical illustration (Section 6); and the worked example (Section 7).

4.5. Half-Synthetic Design

To check that the results do not depend on the parametric noise model, the half-synthetic design starts from the satellite-derived log PM2.5 of 289 cities in 2000–2021. It

keeps the estimated city and year components and multiplies each city's trend-and-residual path, and each province's path of year shocks, by independent Rademacher signs. Known programme effects are then added. With all signs equal to one, the original series is reproduced exactly. The design keeps what a parametric model may miss: each city's own trend and residual path, with its serial dependence, heteroskedasticity and non-normality, and the year shocks common to the cities of a province. The random signs generate new untreated panels and remove two features of the observed data: any alignment between a city's deviations and its designation dates, so that effects the real programmes may have had become noise with mean zero, and any correlation between cities beyond the shared province-by-year shocks. The transformed paths are plausible untreated outcomes if each city's deviation path is as likely as its mirror image, which requires the paths to be centred and roughly symmetric, and if the city and year components are well estimated. Because the expected bias of a linear estimator does not depend on the noise, the design mainly tests inference, that is, whether false-positive rates and coverage survive realistic noise. Its biases differ from those of the main grid mainly because the shorter panel has different weights, and partly because of Monte Carlo error. Selection on trends is removed by construction and is examined in separate designs (Section 5.6). We run the literature-scale reference design for all six focal programmes (300 replications with a non-zero focal effect and 500 with a zero effect) and three main-scale designs per focal programme.

4.6. Estimators

Table 1 lists the fourteen estimators in the main comparison and the applied practice each represents. Standard errors are clustered by city, with province clustering as a robustness check, and tests use t critical values. For the heterogeneity-robust estimators, we use the overall ATT aggregate, and for the event study, the average of post-designation coefficients. For the event study and the two Callaway–Sant'Anna estimators, we also record a joint test of the pre-designation coefficients, which serves as the pretrend test. Six further estimators are reported in Appendix C: the dynamic estimator of de Chaisemartin and D'Haultfœuille [31], the extended TWFE of Wooldridge [35], two variants that condition on other programmes' treatment status, and two spatial-spillover adjustments. The first four were run on a subset of designs because of their computational cost. Deletion designs cannot be estimated when fewer than five treated cities remain, and we report these failures separately rather than dropping them.

Table 1. Estimators and the practices they represent.

Estimator	What It Does	Applied Practice
TWFE	Static TWFE with a post-designation dummy for the focal programme	Baseline DID
Event study	TWFE with leads and lags; averages post-designation coefficients	Dynamic DID with a pretrend check
CS-never	Callaway and Sant'Anna [9], never-treated comparisons	Heterogeneity-robust DID
CS-not-yet	Callaway and Sant'Anna [9], not-yet-treated comparisons	Heterogeneity-robust DID
SA	Sun and Abraham [10] interaction-weighted estimator	Heterogeneity-robust DID
BJS	Borusyak et al. [11] imputation estimator	Heterogeneity-robust DID

Table 1. *Cont.*

Estimator	What It Does	Applied Practice
TWFE + dummies	Adds a post-designation dummy for each other programme	Controlling for other policies
TWFE + dynamic dummies	Adds event-time dummies for each other programme	Richer control for other policies
TWFE + cohort dummies	Adds cohort-by-post dummies for each other programme	Richer control for other policies
Drop-overlapped treated	Deletes focal-treated cities that overlap another programme; comparison group unchanged	Deleting other pilot cities
Stacked, clean controls	Stacked DID whose comparison cities host no other programme	Cleaning the comparison group
Clean both arms	Deletes cities hosting any other programme from both groups	Deleting other pilot cities from the whole sample
Pre-overlap window	Drops every city-year in which another programme is active, in treated and comparison cities alike	Restricting the sample window
Interaction model	TWFE with focal × other-programme interactions; reports the focal effect averaged over the observed overlap	Modelling the overlap

4.7. Performance Measures

We report four measures. Relative bias is the mean estimate minus ATT_f, divided by ATT_f, in designs with a non-zero focal effect. Coverage is the share of nominal 95% confidence intervals that contain ATT_f. The false-positive rate is the share of replications rejecting $\beta_f = 0$ at the 5% level in designs where the focal effect is zero but co-located effects are present. Estimability records whether an estimator can be computed and how many treated cities it retains. We summarise the reference design by the median and interquartile range across the six focal programmes. With 1000 replications, the Monte Carlo standard error of a 5% rejection rate is 0.7 percentage points.

In some designs, the city-specific effect paths are normalised by a random denominator whose theoretical support includes zero; the probability that a draw falls beyond zero is below 3×10^{-14}. Population means of those paths are then not defined, even though no realised draw comes close to zero. For this reason, we summarise designs with medians, coverage and rejection rates, which are well defined, and treat mean biases as averages over the registered replications (Appendix B).

5. Results

The results rest on different parts of the design, and we distinguish three tiers. The first follows from the treatment matrix alone: how much the programmes overlap, how many treated and comparison cities each deletion design retains (Appendix Table A10), how strongly each estimator responds to each co-located programme, as measured by its contamination weights, and which practices set these weights to zero (Section 5.5). These results hold whatever the programmes' effects. The second tier, the bias, also depends on the assumed effects of the co-located programmes. It is linear in those effects: it grows with their size, reverses with their sign, partly cancels when their signs differ and vanishes when they are zero. The headline biases assume that every co-located programme has the focal effect. Appendix Table A8 reports how they change with the ratio of co-located to focal

effects, their timing and their alignment with the focal rollout. Section 5.5 and Appendix Table A9, which gives the range of contamination under alternative assumptions about the co-located effects, show how they change under other sizes and signs. The third tier, false-positive rates and coverage, depends in addition on the noise in the outcome, through Equation (4), and therefore on the PM2.5 calibration and the effect scale.

Table 2 reports simulation performance in the reference design on both effect scales: relative bias, coverage and the false-positive rate. Each entry is the median across the six focal programmes; Appendix Table A4 gives the interquartile ranges, and Table A5 reports selected estimators for each programme separately.

Table 2. Performance in the reference design (medians across six focal programmes).

Estimator	Main: Relative Bias	Main: Coverage	Main: False-Positive Rate	Literature: Relative Bias	Literature: Coverage	Literature: False-Positive Rate	Pilots Estimable
TWFE	120%	79%	18.2%	118%	0%	99.7%	6
Event study	100%	86%	14.3%	97%	9%	89.7%	6
CS-never	78%	86%	12.8%	79%	25%	73.5%	6
CS-not-yet	78%	88%	11.2%	78%	27%	71.3%	6
SA	78%	85%	14.0%	79%	23%	74.9%	6
BJS	139%	76%	21.7%	138%	0%	99.8%	6
TWFE + dummies	−54%	92%	7.8%	−55%	39%	57.7%	6
TWFE + dynamic dummies	−62%	91%	9.1%	−66%	29%	69.1%	6
TWFE + cohort dummies	−50%	91%	8.2%	−48%	51%	50.0%	6
Drop-overlapped treated	−119%	85%	15.8%	−104%	54%	52.8%	4
Stacked, clean controls	89%	84%	22.1%	80%	1%	99.8%	6
Clean both arms	−4%	88%	12.1%	13%	89%	12.4%	4
Pre-overlap window	−18%	91%	11.8%	−16%	87%	10.2%	5
Interaction model	−26%	92%	8.8%	−30%	85%	8.6%	6

Notes: Reference design with dynamic effects (step 0.15), alignment $c = 0.4$ and co-located effects equal in size to the focal effect. Main scale: focal effect 0.2σ (0.57%), 500 replications with a non-zero and 1000 with a zero focal effect per programme. Literature scale: focal effect −3.45%; 500 and 1000 replications. Relative bias and coverage use designs with a non-zero focal effect; the false-positive rate is the rejection rate of a 5% test in designs with a zero focal effect. Standard errors clustered by city. Coverage and false-positive rates are computed over estimable replications; the last column counts focal programmes for which the estimator could be computed. Appendix Table A10 reports the treated and comparison cities retained by each deletion design for each focal programme. Main-scale columns are pre-specified; literature-scale columns are exploratory (Section 4.4).

5.1. Ignoring the Co-Located Programmes

Estimators that ignore the co-located programmes overstate the focal effect substantially. On the main scale, the median relative bias is 120% for TWFE, 100% for the event study, 78% for the two CS estimators and SA, and 139% for BJS. The literature scale gives almost the same relative biases, as Section 3.4 predicts, because bias and focal effect are scaled together.

What changes with the scale is inference. On the main scale, these estimators reject a true null in 11–22% of samples (medians across programmes). On the literature scale, they reject it in 71–100% of samples, and their 95% intervals cover the true effect in, at most,

27% of samples. A researcher evaluating a pilot with no effect, in a city system where the other pilots are as effective as the benchmark, would usually report a significant reduction in PM2.5. Sponge cities are the exception (Appendix Table A5).

The heterogeneity-robust estimators improve on TWFE only modestly, and BJS does worse. Equation (2) explains why. The own-effect weighting term, which these estimators were designed to remove, is small: in the reference design, its median for TWFE across the six programmes is 0.5% of the focal effect and its largest absolute value is 4.3%, and across all designs without interactions, it never exceeds 11%. Almost all the bias is contamination, which the robust estimators leave in place. BJS, which uses every observation not treated by the focal programme to impute counterfactuals, carries the largest contamination weights (Section 5.5).

The bias varies across programmes. For TWFE, it is largest for cross-border e-commerce zones (197%) and e-commerce demonstration cities (171%), and close to zero for sponge cities (7%). For sponge cities, contamination from the co-located programmes' average effects is offset by contamination from their heterogeneity. The same heterogeneity drives the large negative bias of TWFE with dummies for this programme (Appendix Table A5).

5.2. Controlling for the Co-Located Programmes

Adding dummies for the nine co-located programmes reverses the sign of the bias. In the reference design, the median estimate understates the focal effect by 54% on the main scale and 55% on the literature scale. Dynamic and cohort dummies do no better, at −62% and −50% on the main scale and −66% and −48% on the literature scale. On the main scale, their false-positive rates are close to nominal, at 7.8–9.1%, because the contamination is small relative to their standard errors. On the literature scale, they rise to 50–69%.

The grid shows when the dummies work. When all programmes have homogeneous effects, TWFE with dummies is essentially unbiased whatever the size of the co-located effects: its relative bias lies between −4% and 2% for $r = 0.5$, 1 and 2, while that of TWFE rises from 92% to 331%. In the FWL decomposition of Equation (2), contamination from the co-located programmes accounts for a median 86% of the absolute bias of TWFE with dummies across the main grid and 98% in the 36 literature-scale cells for low-carbon and e-commerce cities.

The dummies fail when the co-located effects vary with the focal rollout, and the degree of alignment c decides which specification fails. On the literature scale, with low-carbon or e-commerce cities as the focal programme, TWFE with dummies rejects a true null in 5–9% of samples when $c = 0$, in 23–99% when $c = 0.4$ and in 87–100% when $c = 0.8$, depending on r. TWFE without dummies shows the opposite pattern. It rejects in 96–100% of samples when $c = 0$ but in 0–81% when $c = 0.8$, because the contamination from the co-located programmes' average effects partly offsets that from their heterogeneity. Which specification looks safer, therefore, depends on c, which researchers cannot observe.

5.3. Deleting Observations and Modelling the Overlap

How observations are deleted matters. Deleting overlapped treated cities while keeping the comparison group does not help; it flips the sign of the bias to −119% on the main scale and −104% on the literature scale. The treated cities that remain carry fewer co-located programmes than the comparison cities, so the contamination changes sign. Stacking with clean controls has the mirror-image problem. Treated cities keep their co-located programmes while the comparison cities lose theirs. The bias (89% and 80%) is similar to that of the CS estimators, and on the literature scale, the false-positive rate is 99.8%.

Cleaning both arms works as Section 3.3 predicts. Its contamination weights are zero, and the median relative bias is −4% on the main scale and 13% on the literature scale; the re-

maining deviation reflects partly the change in estimand to cities without other programmes and partly Monte Carlo noise, which is large with so few treated cities. The false-positive rate of about 12% exceeds the nominal level because few treated cities remain. Indeed, very few remain: between 2 and 16 treated cities per programme in the 2000–2022 panel, too few to estimate the e-commerce and sponge-city effects at all. Dropping every city-year in which another programme is active gives similar results, with biases of −18% and −16% and false-positive rates of 12% and 10%, and keeps more treated cities (4 to 60). It estimates the effect only for the years before other programmes arrive, however, and it cannot be computed for sponge cities. Appendix Table A10 reports the treated and comparison cities that each deletion design retains. Cleaning both arms also cuts the comparison group to the 59 cities that host no programme at all. The loss of support costs precision, and with few treated clusters the cluster-robust tests over-reject. It also changes the target: the cleaned designs measure the effect in a small group of cities, or years, without other programmes, and Section 2.2 shows that cities without other programmes differ systematically from those that host several. A difference between full-sample and cleaned estimates therefore combines the removal of contamination with a change in estimand and a loss of precision.

The interaction model is the only full-sample method whose false-positive rate stays close to the nominal level on both scales: 8.8% and 8.6% in the reference design, with coverage of 92% and 85%. Across the 132 zero-effect cells of the main grid, its false-positive rate ranges from 5.2% to 16.3%, against 3.9% to 45.1% for TWFE with dummies. It still understates the focal effect by 26–30%, because it treats each combination of programmes as having a constant effect while the simulated effects vary over time and across cities.

We also ran the dynamic estimator of de Chaisemartin and D'Haultfœuille [31], in its standard form and in the several-treatments form of de Chaisemartin and D'Haultfœuille [12], which compares switching cities only with cities that share their history of other programmes (Appendix Table A7). The standard estimator has a false-positive rate of 6.5% on the main scale but 44.8% on the literature scale, where it was run for low-carbon and e-commerce cities with co-located effects one and two times the focal effect. Measured against its own estimand, it carries contamination of 36% and 66%. The several-treatments form removes most of this contamination on the literature scale, with a bias of 4% relative to its own estimand and a false-positive rate of 17% (12% with province clustering). It does so by discarding most of the data: only 6.5% of switching cities have a valid comparison on the literature scale and 14% on the main scale, and on the main scale its bias relative to its own estimand is 29%.

No method in Table 2 recovers the full-sample average effect free of contamination. The methods with zero contamination weights change the estimand and lose most treated cities. The methods that keep the full sample carry contamination of between −65% and +137% of the focal effect on the main scale; the interaction model carries the least, at −22%. Appendix Table A6 splits each method's bias into the two terms of Equation (2).

5.4. Why False Positives Explode, and Why Pretrend Tests Miss Them

Figure 2 plots the false-positive rate of every estimator in every design against the ratio of absolute bias to the mean estimated standard error. On the literature scale, the points lie on the curve implied by Equation (4), with a correlation of 0.997: the over-rejection is almost fully accounted for by the bias relative to the standard error. With city-clustered standard errors, the median ratio is 2.6, and in a quarter of the cells, it exceeds 4.8, where rejection is almost certain. On the main scale, the ratios are smaller, with a median of 0.31, and the standard errors are slightly too small (the Monte Carlo standard deviation exceeds the mean estimated standard error by 6% at the median), so both sources contribute.

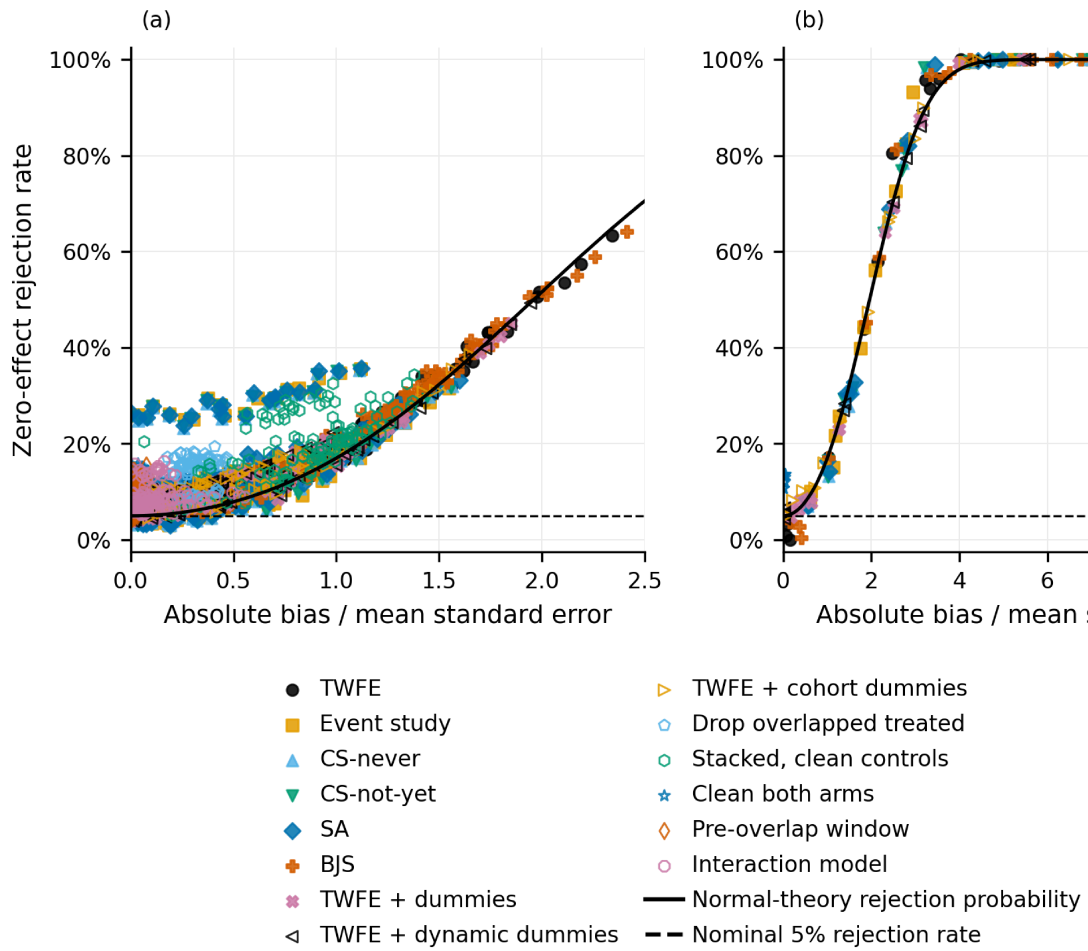


Figure 2. False-positive rates and the ratio of bias to standard error. Each point is one estimator in one design with a zero focal effect, using city-clustered standard errors; marker shapes distinguish estimators. The curve is Equation (4). (**a**) Main scale, all 132 zero-effect cells of the main grid; (**b**) literature scale, the 18 zero-effect cells for low-carbon and e-commerce cities. The dashed line marks 5%. The normal curve is a reference; the reported tests use cluster-robust standard errors and t critical values.

Figure 3 compares contaminated designs with the benchmark that has no co-located effects. On the main scale, pretrend tests pass at almost the same rate in both: 89–94% in contaminated designs against 91–93% in the benchmark. At this effect size, the tests carry almost no information about contamination. On the literature scale, they fail more often in contaminated designs, with pass rates of 76–83% against about 95%, but passing offers no protection. Among contaminated designs that pass, the false-positive rate is 82–85%, against 5% in the benchmark, and it is higher than among those that fail (68–75%). Contamination enters mostly after designation, when co-located programmes arrive or their effects grow, so a test on pre-designation coefficients sees only the part that arrives before designation (see also [36]).

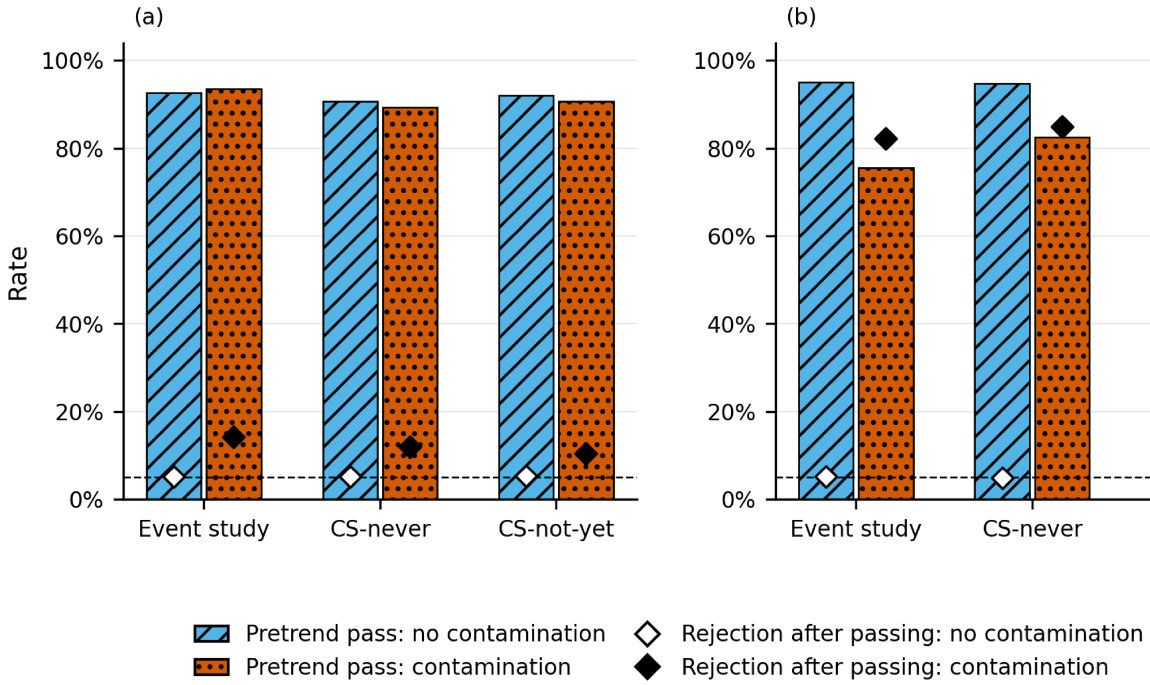


Figure 3. Pretrend tests in clean and contaminated designs with a zero focal effect. Bars show the share of replications that pass the pretrend test (hatched: no contamination; dotted: contamination); diamonds show the false-positive rate among replications that pass (open: no contamination; filled: contamination). Clean designs have no co-located effects; contaminated designs are the reference design. (**a**) Main scale, six focal programmes; (**b**) literature scale, low-carbon and e-commerce cities. Medians across focal programmes; city-clustered standard errors. The event study uses a cluster Wald test of the leads; the CS estimators use their aggregated lead test. The dashed line marks 5%.

5.5. Contamination Weights (Exploratory)

Table 3 reports the contamination weights of Equation (3), summed over the nine co-located programmes, for three unit effect paths. Each entry is the range across the six focal programmes. A sum of 1.5 means that if every co-located programme changed PM2.5 by 1% on average, the estimate for the focal programme would move by about 1.5%.

The table makes the mechanisms of Section 3.3 visible. Estimators that ignore co-located programmes have large positive sums for every path. Adding dummies sets the weights to zero for exactly the paths the dummies span: static dummies remove static effects, dynamic dummies remove common dynamic paths, cohort dummies remove cohort-specific effects. Every other path leaves weights of either sign. Cleaning both arms or truncating the window sets all weights to zero, whereas deleting overlapped treated cities alone produces large negative weights.

The weights reproduce the simulated bias. In the 48 cells in which expected effect paths are defined analytically (42 main-grid and 6 half-synthetic cells with homogeneous or own-cohort effects), the bias predicted from the weights matches the Monte Carlo bias with an R^2 of 0.997 for TWFE, 0.993 for CS-never and SA, 0.997 for BJS and 0.994 for the event study, and the largest discrepancy is 2.7 Monte Carlo standard errors. For TWFE with dummies, the R^2 is 0.74, because its bias varies little across these cells, but the largest discrepancy is 2.2 standard errors. Applying each estimator to the average simulated signal of the registered replications reproduces the Monte Carlo mean with an R^2 of at least 0.98 for TWFE, the event study, the CS, SA and BJS estimators and the three dummy specifications, and 0.94 for the interaction model. For linear estimators, this second check is close to an identity and confirms the implementation rather than providing independent validation.

Table 3. Summed contamination weights by estimator and effect path (range across six focal programmes).

Estimator	Static Effects	Cohort-Specific Effects	Dynamic Effects
TWFE	1.23 to 2.28	1.00 to 1.64	1.26 to 2.67
Event study	0.42 to 1.76	0.57 to 1.22	0.88 to 2.20
CS-never	0.26 to 1.75	0.57 to 1.20	0.77 to 2.23
CS-not-yet	0.23 to 1.75	0.50 to 1.19	0.71 to 2.22
SA	0.26 to 1.75	0.57 to 1.20	0.77 to 2.23
BJS	1.61 to 2.48	1.03 to 1.72	1.90 to 2.76
TWFE + dummies	0	−0.50 to 0.16	−0.08 to 0.58
TWFE + dynamic dummies	0	−0.35 to 0.33	0
TWFE + cohort dummies	0	0	−0.03 to 0.56
Drop-overlapped treated	−1.11 to −0.63	−1.38 to −0.93	−1.18 to −0.71
Stacked, clean controls	0.51 to 1.57	0.69 to 1.52	0.83 to 1.79
Clean both arms	0	0	0
Pre-overlap window	0	0	0
Interaction model	0	0.24 to 0.99	−0.50 to −0.28

Notes: Each path is normalised to a mean of one over the co-located programme's own treated city-years. Weights are computed by applying the estimator to noise-free outcomes equal to each co-located programme's path, in the treatment matrix of the reference design. Drop-overlapped treated and clean both arms cover four focal programmes; pre-overlap window five. Entries of 0 are below 0.01 in absolute value.

Because the weights depend only on the design, they also show how the results change when the co-located effects depart from the reference assumption that all nine equal the focal effect (Appendix Table A9; literature scale, medians across the six focal programmes). If only the two programmes that overlap most with the focal programme have effects, the estimators that ignore co-located programmes carry contamination of 18–40%, and the false-positive rates implied by Equation (4) are 10–30%; with three such programmes, the figures are 33–66% and 22–61%. Drawing each programme's effect independently as half, once or twice the focal effect puts their contamination between 60% and 213% in 90% of draws for the median focal programme. The sign matters as much as the size. If the co-located programmes work in the opposite direction, every weight acts in reverse: TWFE and BJS then report an effect of the wrong sign, the event study an effect close to zero, and the CS and SA estimators understate the effect by 77–78%, while adding dummies overstates it by 47–65%. If half of the programmes raise the outcome and half lower it, contamination is centred near zero, but its 5th–95th percentile range is still about ±45% for TWFE and ±16% for the interaction model. Re-simulating 16 of these cases with the full pipeline, 200 replications each, reproduced the predicted biases to within 0.0014 log points. If the co-located programmes have no effect, there is no contamination at all. The weights fix how strongly each estimator responds to the co-located effects, and the assumed effects fix the direction and size of the response; the reference design, in which all nine programmes share the sign and size of the focal effect, is one scenario among these rather than a prediction.

Simple summaries cannot replace the calculation. Three overlap indices, namely the overlap share of treated city-years, the largest correlation between the focal and other treatment indicators after removing city and year effects, and the sum of absolute FWL weights, predict the bias for a held-out programme family only partly. The leave-one-family-

out R^2 is 0.57–0.62 for the event study, CS and SA, 0.28 for TWFE with dummies, 0.20 for BJS and −0.03 for TWFE. Summary statistics of the weights do worse than a constant out of sample, with leave-one-family-out R^2 between −5.7 and −1.5 for these estimators. The weights are informative about the design at hand, not as a universal rule.

5.6. Robustness

Real outcome paths. The half-synthetic design keeps the satellite-derived PM2.5 paths and reproduces the main findings. In the literature-scale reference design, the median relative bias is 113% for TWFE, 73% for CS-never and 130% for BJS, with false-positive rates of 99.9%, 76.9% and 100%. TWFE with dummies has a bias of −51% and a false-positive rate of 59.2%. Cleaning both arms, the pre-overlap window and the interaction model have false-positive rates of 9.4%, with biases of 11%, −14% and −20%. Because the design keeps the observed noise rather than a parametric model of it, these figures speak mainly to inference: the contaminated estimators still over-reject heavily, while the clean designs and the interaction model do so only mildly. As the design removes correlation between cities beyond the province-by-year shocks, these rates may understate over-rejection with city-clustered standard errors if such correlation is present in the data.

Clustering. Clustering by province lowers the false-positive rates of the CS estimators, from 12.8% to 7.0% on the main scale and from 73.5% to 46.6% on the literature scale for CS-never. It barely changes those of TWFE (99.7% to 98.4%) and BJS (99.8% to 98.5%) on the literature scale. Province clustering is advisable, since pilot programmes cluster geographically, but it does not remove the problem.

Spillovers, anticipation, selection and serial correlation. Spatial spillovers leave the results essentially unchanged: relative biases of 126% for TWFE, 86% for CS-never and −53% for TWFE with dummies. One-year anticipation lowers the bias of CS-never to 18% but not that of TWFE (109%) or TWFE with dummies (−64%). Varying the autocorrelation between 0 and 0.8 changes little. Selection of cities on their underlying trends adds bias of its own. Under unconditional selection, the relative bias of the clean-both-arms design reaches 67%, so deletion is no remedy when selection is present.

Coding of the treatment matrix. Overlap persists under alternative codings. Dropping the 64 designations awaiting final confirmation leaves 160 cities with two or more programmes; counting only whole-city designations leaves 168, and treating unverified later national batches as treated raises the count to 297. We recompute the overlap diagnostics under each coding but, for reasons of computing time, do not repeat the simulations.

Interactions between programmes. When the focal effect depends on how many other programmes a city hosts, "the effect of the focal programme" is no longer a single number. Estimators then differ in which mix of overlap configurations they average over, and none recover the effect averaged over the observed configurations in both interaction designs (Appendix Table A8).

6. Empirical Illustration: Satellite-Derived PM2.5 (Exploratory)

We now apply the estimators to satellite-derived log PM2.5 for 289 cities in 2000–2021 (6358 city-years), using the treatment matrix restricted to these cities and years (Section 4.1). The exercise shows how much conclusions depend on the handling of overlap and how large contamination could be under stated assumptions about the co-located effects. It does not identify the programmes' effects, since selection and other policies, such as the 2013 Air Pollution Prevention and Control Action Plan, are not modelled. Table 4 reports the results.

Table 4. Estimates for satellite-derived log PM2.5 and the contamination that co-located programmes could induce.

Estimator	Low-Carbon	E-Commerce	New-Energy Vehicles	Sponge	Civilised	Cross-Border
TWFE	−1.91 *	0.20	−0.64	−3.02 *	−0.84	−1.92 *
Event study	−1.74 *	−0.70	−1.11	−3.42 *	−1.49 *	−1.79 *
CS-not-yet	−1.85 *	−0.82	−0.89	−3.28 *	−2.15 *	−1.93 *
SA	−1.85 *	−0.79	−0.85	−3.26 *	−2.21 *	−1.93 *
BJS	−2.15 *	0.06	−0.76	−3.09 *	−1.15	−2.07 *
TWFE + dummies	−1.71	1.69	−0.16	−2.30	−0.62	−1.09
Clean both arms	−3.49	−2.15	1.08	n.a.	1.90	4.83 *
Pre-overlap window	−1.42	−1.56	−0.11	n.a.	0.59	4.54 *
Interaction model	1.21	4.24 *	0.91	0.58	0.11	0.97
Contamination if each co-located programme had an effect of −3.45%						
TWFE, static paths	−5.4	−7.6	−5.4	−5.8	−4.1	−7.7
TWFE, dynamic paths	−6.1	−8.4	−6.2	−7.7	−4.2	−10.4
CS-not-yet, static paths	−3.7	−5.8	−3.2	−1.5	−4.7	−0.8
CS-not-yet, dynamic paths	−4.9	−7.3	−4.4	−3.1	−5.4	−2.6

Notes: Estimates are 100 × coefficients on log PM2.5, approximately percent. The exact percentage change, $100 \times (\exp(b) - 1)$, differs from the entries by at most 0.12 percentage points in the upper panel and by at most 0.52 points in the contamination rows (−10.4 corresponds to −9.9%). * $p < 0.05$ with city-clustered standard errors. With province clustering, the cross-border estimates of TWFE, CS-not-yet, BJS, clean both arms and the pre-overlap window lose significance at 5%, and the TWFE + dummies estimate for e-commerce becomes significant ($p = 0.04$). n.a.: fewer than five treated cities remain (2 and 4 for sponge cities). TWFE, the event study, CS-not-yet, SA, BJS, TWFE + dummies and the interaction model use the full PM2.5 panel: 68, 65, 82, 27, 144 and 95 treated cities for the six programmes in column order, and 221, 224, 207, 262, 145 and 194 comparison cities. Treated cities retained by clean both arms: low-carbon 6, e-commerce 5, new-energy vehicles 8, civilised 20, cross-border 7; by the pre-overlap window: 26, 11, 38, 58 and 7. Comparison cities (not designated for the focal programme in the retained sample): 66 for every programme when both arms are cleaned, and 263, 278, 251, 231 and 282 in the pre-overlap window (Appendix Table A10). The last four rows apply the contamination weights computed for this sample to an average effect of −3.45% for every co-located programme. Retained counts refer to the 289-city panel for 2000–2021 and therefore differ from those for the full panel in Sections 5.3 and 7. The rows are not directly comparable, because their estimands differ: the full-sample estimators average over all treated city-years with estimator-specific weights, the interaction model averages the focal effect over the observed overlap, clean both arms refers to cities that never host another programme, and the pre-overlap window refers to the years before another programme arrives. Differences across rows can therefore reflect contamination, sample support and the estimand.

The standard and heterogeneity-robust estimators agree with one another. All five find significant reductions of 1.7–2.2% for low-carbon cities, 3.0–3.4% for sponge cities and 1.8–2.1% for cross-border e-commerce zones. The event-study, CS and SA estimates are also significant for civilised cities, at −1.5% to −2.2%. A reader of any single study using one of these estimators would find a credible-looking reduction.

The bottom rows show why this agreement is not reassuring. If each of the nine co-located programmes reduced PM2.5 by 3.45% on average, contamination alone would shift the TWFE estimate by 0.041–0.104 log points (about 4.0–9.9%) across the six programmes

and the CS-not-yet estimate by 0.008–0.073 log points (0.8–7.0%). For low-carbon cities, the TWFE contamination would be 0.054–0.061 log points (5.3–5.9%), about three times the estimate itself. The estimates therefore lie well within the range that contamination could produce if co-located programmes are as effective as the published benchmark. If those programmes had no effect, the contamination would vanish. If they raised PM2.5, the same weights would shift the estimates upwards by the same amounts. The data alone cannot tell these cases apart. The bottom rows therefore show how far the estimates could move under stated assumptions; they are not estimates of the contamination actually present.

Estimators that limit contamination give a different, if imprecise, picture. Cleaning both arms and truncating the window remove the significant reduction for low-carbon cities and reverse the sign for civilised cities and cross-border zones; for the latter, both give significant increases of 4.5–4.8%, although each rests on only seven treated cities. The interaction model gives positive point estimates for all six programmes, significant for e-commerce cities (4.2%). TWFE with dummies yields no estimate that is significant with city-clustered standard errors.

The cross-border reversal should not be read as the corrected effect. Both designs rest on the same seven cities, all ordinary prefecture-level cities designated in 2019 or 2020 and therefore observed for at most three years after designation. They are, by construction, cities that hosted no other programme, and Section 2.2 shows that such cities differ systematically from those that host several. The two estimates therefore describe a small and selected group rather than the 95 cross-border cities in this panel, and neither remains significant with province clustering. The simulations point the same way: with so few treated cities, these designs rejected a true null in 10–12% of samples, more than twice the nominal rate (Table 2). The reversal may reflect the removal of contamination, the change in sample and estimand, or chance, and these data cannot tell them apart. The same caution applies to civilised cities, where the clean-both-arms estimate of 1.90 rests on 20 treated cities and is not significant.

We do not read these numbers as programme effects. For all six programmes, defensible estimators disagree on the sign of the effect, its significance or both, and the overlap structure alone can generate differences in this size.

7. Implications for Evaluation Practice

For evaluators, the results suggest a short protocol for any DID evaluation of a programme that shares its units with other programmes. The first three steps use only the treatment matrix, so they can be written into a pre-analysis plan before any outcome data are examined.

1. **Map the overlap.** List the national and provincial programmes that plausibly affect the outcome and code them at the same level and with the same rules as the focal programme. Report the share of the focal programme's treated city-years that host other programmes, the average number of such programmes, and the largest pairwise overlaps, as in Figure 1 and Table A1.
2. **Compute the contamination weights of the chosen estimator.** For each co-located programme, run the estimator on synthetic outcomes equal to the programme's treatment indicator (static path), to its normalised years since designation (dynamic path) and to cohort-specific paths. This takes seconds, needs no outcome data, and gives the weights of Equation (3).
3. **Translate the weights into a bias range.** Multiply the weights by plausible effects of the co-located programmes, taken from published estimates or from a range around the focal estimate. If the implied contamination is comparable to the estimate, the estimate says little about the focal programme alone. Because the weights are signed,

consider effects of both signs: co-located programmes that work against the focal programme reverse the direction of the bias (Section 5.5). The result is a range under stated assumptions, not an estimate of the bias actually present; Appendix Table A9 shows such ranges for the reference design.

4. **Do not treat dummies or pretrend tests as evidence of no contamination.** Dummies remove only the effect paths they span, and pretrend tests cannot see contamination that arrives after designation.
5. **Report a design that is clean in both arms**, or a pre-overlap window, wherever enough treated units remain. Report the numbers of treated and comparison units retained and the estimand the design identifies. Do not clean only one arm. The several-treatments estimator of de Chaisemartin and D'Haultfœuille [12] automates this cleaning; report how many switching cities it retains.
6. **Report an interaction model** alongside the main estimate. In our designs, it was the only full-sample method whose false-positive rate stayed close to the nominal level on both effect scales.
7. **Cluster by province** as a robustness check, since programmes cluster geographically.
8. **State the estimand when programmes may interact:** the effect in cities without other programmes, or the effect averaged over the combinations actually observed.

The clean designs of step 5 will often be unavailable. Among our six focal programmes, cleaning both arms leaves at least five treated cities for four programmes and the pre-overlap window for five; for sponge cities, neither design can be estimated. Where they can be estimated, they rest on 5–16 and 11–60 treated cities, respectively; cleaning both arms leaves only the 59 cities without any programme as comparisons (Appendix Table A10), and in the simulations they rejected a true null in 10–12% of samples. A researcher following the protocol should therefore expect that, for many programmes, no estimator is both free of contamination and precise. Steps 1–3 remain available in every case, because they need only the treatment matrix, and so does the interaction model. Where no clean design can be estimated, the appropriate report is the full-sample estimate together with the range of contamination implied by the weights, and a conclusion that the data cannot separate the focal programme from its co-located programmes.

A worked example. Table 5 shows the output of the first three steps for the low-carbon city pilot. It uses the treatment matrix of Section 2 and no outcome data, and the calculation takes about two seconds on a laptop. Of the low-carbon pilot's treated city-years (70 treated cities), 82% coincide with at least one other programme. Requiring both arms to be free of other programmes would leave 5 treated cities, and censoring each city's series when another programme arrives would leave 26. The weights show why the choice of estimator matters. For TWFE, the summed weights are 1.64 for static effects, 1.86 for dynamic effects and 1.17 for cohort-specific effects: if every co-located programme had an average effect equal to the focal one, TWFE would overstate the focal effect by 117–186%. Adding dummies for the other programmes removes the static weights exactly but leaves 0.09 for dynamic and −0.23 for cohort-specific effects, so the size and even the sign of the remaining bias depend on how the other programmes' effects evolve. An evaluator of this pilot would therefore report the weights and the implied range of bias next to the estimate, add the design that is clean in both arms with its five treated cities, and not present the dummy specification as evidence that other policies have been controlled for.

Table 5. Contamination weights for the low-carbon city pilot, summed over the nine co-located programmes.

Estimator	Static Effects	Dynamic Effects	Cohort-Specific Effects
TWFE	1.64	1.86	1.17
TWFE + dummies	0.00	0.09	−0.23

Notes: Computed by the function overlap_diagnostic() from the treatment matrix alone. Each path is normalised to a mean of one over the co-located programme's own treated city-years, so an entry is the relative bias that results if every co-located programme has the focal programme's average effect. For TWFE + dummies with cohort-specific effects, the absolute weights sum to 0.47. Support: 70 treated cities; 82.3% of treated city-years overlap another programme; 5 treated and 59 comparison cities remain when both arms are cleaned, and 26 treated and 271 comparison cities in the pre-overlap window.

The function overlap_diagnostic() in the replication package carries out steps 1 to 3 for every estimator in Table 1. It takes a city–year treatment matrix, returns the overlap and support statistics and the weights by programme and effect path, and converts user-supplied effect sizes into an implied bias.

8. Conclusions

Pilot programmes are one of China's main instruments of policy learning, and DID evaluations of them are among the most common designs in Chinese applied economics. This paper shows that the overlap of programmes in the same cities is a first-order threat to those evaluations. In our calibrated designs, it matters far more than the heterogeneity problem that recent DID methods address. Estimators that ignore co-located programmes overstate effects when those programmes work in the same direction as the focal one, and understate or reverse them when they work in the opposite direction. Adding dummies for those programmes reverses the bias whenever their effects vary with the focal rollout. Pretrend tests rarely flag the contamination and passing them offers no protection, and deleting overlapped cities works only when both arms are cleaned, at the cost of most treated cities. Because the contamination is a weighted sum of the other programmes' effects with weights fixed by the design, its range under assumed effects can be computed before estimation, and we suggest that it be reported. These results differ in scope. The contamination weights are fixed by the treatment matrix and the estimator and hold whatever the programmes' effects. The biases also depend on the assumed effects of the co-located programmes, and the false-positive rates depend in addition on the noise in the outcome, so the rates reported here describe the simulated scenarios and are not predictions for any particular evaluation. The empirical illustration makes the same point with observed data: estimates for the same programme differ in sign or significance across defensible estimators, and the contamination that the published benchmark implies for TWFE is larger than every TWFE estimate. It shows how fragile such estimates can be, not the direction or size of any programme's effect.

The stakes go beyond method. Pilots are how China decides which environmental and low-carbon measures to extend nationwide, and DID estimates of this kind are meant to inform that choice. An evaluation that credits one pilot with the effects of its co-located programmes can favour a measure that does little and obscure one that works. Reporting contamination weights next to such estimates is a cheap safeguard for evidence-based sustainability policy.

Several limitations qualify these conclusions. First, our database covers ten programmes and excludes later national batches that we could not verify. True overlap is probably larger, because those batches, other national pilots such as smart cities, provincial programmes and non-pilot policies such as the 2013 Air Pollution Prevention and Control

Action Plan are not coded. Second, the outcome model is calibrated to PM2.5. Relative biases depend on the ratio of co-located to focal effects, their alignment and their paths, and carry over to other outcomes only when these are similar; false-positive rates also depend on each outcome's noise. Third, the literature-scale benchmark rests on three verifiable published estimates, and assuming that every co-located programme is equally effective makes it a stress test rather than a forecast. Fourth, the half-synthetic design assumes that each city's deviation path is as likely as its mirror image, and it removes correlation between cities beyond the province-by-year shocks. If the observed paths are not symmetric, its replications misrepresent the untreated outcomes, and if outcomes are correlated across cities, its false-positive rates may understate over-rejection; the design does not address selection, which the selection designs do. Fifth, we did not evaluate every available remedy. The several-treatments estimator of de Chaisemartin and D'Haultfœuille [12] was run only in the reference designs and with fewer replications. Synthetic DID [37] and covariate-adjusted estimators are also left for future work.

Two extensions seem especially useful. A public overlap atlas of Chinese pilot programmes would let researchers map co-located programmes in minutes rather than weeks. And because contamination weights require only the treatment matrix, journals could ask authors to report them as routinely as pretrend tests.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/su182111300/s1, Table S1: batches of the ten pilot programmes: official sources and coding (file Table_S1_programme_batches.xlsx); Script S1: make_Table_S1.py, which generates Table S1 from the replication package.

Author Contributions: Conceptualisation, C.C. and J.Z.; methodology, C.C.; software, C.C.; formal analysis, C.C.; data curation, C.C.; writing—original draft preparation, C.C.; writing—review and editing, C.C.; visualisation, C.C.; supervision, J.Z., and J.Z. also contributed to the interpretation of the results. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The replication package is available at https://osf.io/qfaps/overview?view_only=cf4ee02edb4b4902a80c8b8485c0ae2b (accessed on 4 October 2026). It contains the city–programme–year treatment table (297 cities, 2000–2022); a record-level table of the 760 designations with their official sources, web addresses and retrieval dates; the batch-level source table on which Supplementary Table S1 is based; area-weighted city-year aggregates of log PM2.5 for 289 cities in 2000–2021 and their decomposition; the simulation and estimation code with locked package versions; the pre-analysis plan; the function overlap_diagnostic() with a worked example; and the summary results from which all tables and figures are reproduced. It does not contain the per-replication simulation output, which is not needed to reproduce the tables and which the code of the main simulation pipelines regenerates from the recorded seeds; the ChinaHighPM2.5 rasters (Zenodo, https://doi.org/10.5281/zenodo.6398971), whose CC BY 4.0 licence also covers the city aggregates; the GADM 4.1 city boundaries used for aggregation, whose licence permits academic use but not redistribution; or copies of the government documents, which are identified by their web addresses. Supplementary Table S1 (Table_S1_programme_batches.xlsx) is generated by the script make_Table_S1.py, supplied with it, from three files in the replication package: frozen_summaries/Table_A1_batches.csv (batch sources), data/public/policy_adoptions_with_sources.csv (the 760 designations) and config/policies_stage2.yml (the units named in each official list).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Programmes and Effect-Size Benchmark

Table A1. Pilot programmes and their overlap, 2000–2022.

Programme	Treated Cities	Treated City-Years	Share of Treated City-Years with Another Programme	Mean Number of Other Programmes
Low-carbon city pilots	70	606	82%	3.2
Innovative city pilots (NDRC)	17	223	91%	4.4
Cross-border e-commerce pilot zones	152	467	93%	3.9
E-commerce demonstration cities	68	651	93%	3.7
New-energy-vehicle promotion cities	85	827	80%	2.8
Sponge city pilots	27	202	91%	3.9
Social credit demonstration cities	53	176	95%	4.6
Energy-saving fiscal demonstration cities	30	304	77%	2.8
National civilised cities	148	1168	79%	2.5
Broadband China demonstration cities	110	881	75%	2.6

Notes: 297 prefecture-level cities, 2000–2022. A city is treated from the year of its official designation if the city or any of its urban districts is designated. Table A3 summarises the batches, issuing authorities and coding of each programme; Supplementary Table S1 lists the official source of each batch, and the replication package documents every record.

Table A2. Published log-PM2.5 DID estimates used for the literature scale.

Study	Programme	Coefficient (Log Points)	Implied Change
Yan et al. [34]	Low-carbon city pilots	−0.1429	−13.32%
Deng and Guo [30]	Low-carbon city pilots	−0.0351	−3.45%
Li et al. [28]	Cultural consumption pilots	−0.0149	−1.48%

Notes: The three studies are those of Chinese city pilots whose published full texts report a main DID coefficient on annual log PM2.5 that we could verify. The median, −0.0351, defines the literature scale. Studies that report levels, emissions or percentages without a verifiable log coefficient were excluded. The benchmark is not a meta-analytic estimate.

Table A3. Programme batches, sources and coding.

Programme	Issuing Authority	Batches: Year (Cities)	Coding Notes, Uncertain and Excluded Cases
Low-carbon city pilots	NDRC	2010 (8); 2012 (26); 2017 (36)	Pilot provinces not extended to their cities; 2012 list from the official reprint by the National Center for Climate Change Strategy and International Cooperation; 15 cities of 2017 approved subject to revised plans (flagged). Listed counties, county-level cities and prefectures not coded.

Table A3. *Cont.*

Programme	Issuing Authority	Batches: Year (Cities)	Coding Notes, Uncertain and Excluded Cases
Innovative city pilots (NDRC)	NDRC	2008 (1); 2010 (16)	NDRC sequence only; later innovative-city lists of the Ministry of Science and Technology not merged. Shenzhen (2008) dated by an official NDRC report of 2010.
Cross-border e-commerce pilot zones	State Council	2015 (1); 2016 (12); 2018 (21); 2019 (22); 2020 (43); 2022 (24; 29)	Seven approvals, two in 2022, dated by year of approval. Xiong'an New Area, autonomous prefectures, prefectures and county-level cities (e.g., Yiwu, Hunchun, Manzhouli) not coded.
E-commerce demonstration cities	NDRC and seven other central agencies	2009 (1); 2011 (22); 2014 (29); 2016 (16)	Changchun–Jilin joint designation mapped to both cities. Shenzhen (2009) dated by an official retrospective report; 2016 notice, posted online in 2017, coded 2016. Yiwu and Wujiaqu not coded.
New-energy-vehicle promotion cities	MOF, MOST, MIIT and NDRC	2013 (62); 2014 (23)	2013–2015 programme only; the 2009–2010 demonstration not merged. City groups expanded to member cities. Qiandongnan Prefecture not coded.
Sponge city pilots	MOF, MOHURD and MWR	2015 (13); 2016 (14)	Official publicity lists of the competitive review; final funding lists not verified (all 27 flagged); 2021 demonstration programme not merged. Qian'an, Gui'an New Area and Xixian New Area not coded.
Social credit demonstration cities	General Offices of NDRC and PBoC	2017 (10); 2019 (15); 2021 (28)	Third batch named demonstration areas; districts mapped to their city. Fourth batch (2023) outside the sample period. County-level cities (e.g., Yiwu, Rongcheng, Kunshan) not coded.
Energy-saving fiscal demonstration cities	MOF and NDRC	2011 (8); 2013 (10); 2014 (12)	2013 and 2014 batches from official publicity lists; final funding lists not verified (22 flagged).
National civilised cities	Central Steering Commission for Spiritual Civilization Construction	2005 (11); 2009 (10); 2011 (23); 2015 (28); 2017 (34); 2020 (42)	First award only; later reviews and withdrawals not coded. Awards to urban districts mapped to their city; awards to counties and county-level cities, numerous in 2017 and 2020, not coded. Seventh cohort outside the sample period.

Table A3. *Cont.*

Programme	Issuing Authority	Batches: Year (Cities)	Coding Notes, Uncertain and Excluded Cases
Broadband China demonstration cities	MIIT and NDRC	2014 (37); 2015 (37); 2016 (36)	Chang–Zhu–Tan cluster expanded to three cities; Jiulongpo and Beibei (2016) mapped to Chongqing; Fuzhou includes Pingtan. Autonomous prefectures and county-level cities not coded.

Notes: NDRC, National Development and Reform Commission; MOF, Ministry of Finance; MOST, Ministry of Science and Technology; MIIT, Ministry of Industry and Information Technology; MOHURD, Ministry of Housing and Urban–Rural Development; MWR, Ministry of Water Resources; PBoC, People's Bank of China. Numbers in parentheses are the cities first designated in each batch; they sum to 760, and the two cross-border approvals of 2022 are shown separately. A batch can name more units than it adds cities, because lists also name counties, county-level cities, new areas, autonomous prefectures and prefectures, which lie outside the 297-city panel, districts of cities already designated, and cities designated in an earlier batch. The 64 flagged designations are dropped in the alternative coding of Section 5.6. Supplementary Table S1 gives, for each batch, the official document, web address, retrieval date, the numbers of units listed and cities coded, and the listed units that add no city.

Appendix B. Simulation Details

Outcome model. Treatment histories are fixed at the observed matrix, $D_{kit} = 1\{t \geq G_{ki} > 0\}$, where G_{ki} is the designation year. The parametric outcome is

$$Y_{it} = \alpha_i + \lambda_t + (b_i + \ell X_i)(t - t_0) + U_{p(i)t} + \varepsilon_{it} + \sum_k \tau_{kit} D_{kit} + \kappa\sigma D_{fit} \sum_{k \neq f} D_{kit} + \omega\sigma \sum_{j \neq i} W_{ij} D_{fjt} + a\sigma\, 1\{t = G_{fi} - 1\},$$

with $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + \eta_{it}$. City effects, year effects, city trends, province-by-year shocks and innovations η_{it} are independent normal draws with standard deviations 0.2842, 0.1914, 0.0047, 0.0493 and 0.0266, and $\rho = 0.3557$. All are estimated from a two-way decomposition of the ChinaHighPM2.5 city panel (289 cities, 2000–2021), with a small-sample correction for the 31 provinces. The scale $\sigma = 0.0284$ is the standard deviation of the idiosyncratic residual after removing city effects, year effects, city trends and province-by-year components; effect sizes are expressed in multiples of it.

Effect paths. Let x_i be the standardised focal designation year, with never-designated cities set to two years after the panel ends. For each co-located programme k, the city modifier is $m_{ki} = c\, x_i + (1 - c^2)^{1/2} z_{ki}$, where z_{ki} is a standardised normal vector orthogonalised on x_i, so that its sample correlation with x_i equals c exactly. Heterogeneous designs use $h_{fi} = 1 + 0.25 x_i$ for the focal programme and $h_{ki} = 1 + 0.4 m_{ki}$ for co-located programmes; homogeneous designs set both to one. Dynamic designs multiply by $q(e) = 0.5 + s \cdot \min(\max(e, 0), 5)$, where e is years since designation. Each path is then divided by its mean over the programme's treated city-years and multiplied by the programme's average effect μ_k.

Interactions, spillovers, anticipation and selection. In the interaction designs, each active co-located programme adds $\kappa\sigma$, with $\kappa = \pm 0.1$, to the focal effect. Spillovers use equal weights W_{ij} over other cities whose centres lie within 200 km, with $\omega = 0.1$. Anticipation adds $a\sigma$, with $a = 0.1$, in the year before designation. The selection designs add a trend of 0.002 log points per year per standard deviation of standardised longitude X_i; the conditional design adjusts for it and the unconditional design does not. The targets exclude spillover and anticipation effects.

Half-synthetic design. The baseline is the grand mean plus the estimated city and year effects plus $u_i\left[\hat{b}_i(t - \bar{t}) + E_{it}\right] + v_p \hat{U}_{pt}$, where u_i and v_p are independent Rademacher

signs for each city and province, $\hat{b}_i$ are the estimated city trends, E_{it} the residuals and $\hat{U}_{pt}$ the province-by-year components. Known programme effects are added to this baseline.

Normalisation of effect paths. For each programme, city-specific deviations are drawn and the resulting path is divided by its realised mean over the programme's treated city-years, so that the programme's average effect equals μ_k exactly in every replication. In 1551 of the 4932 programme-by-cell combinations, the theoretical support of this denominator includes zero. The probability that it falls on the far side of zero is at most 3.0×10^{-14} per draw, with a median of 1.2×10^{-69}. The first absolute moment of the normalised effects, and hence the population mean bias, is then undefined, although every realised replication is numerically stable. We therefore summarise with medians, coverage and rejection rates, which are well defined, and interpret mean biases as averages over the registered replications. In an audit of 322,000 draws across 548 cells, the realised denominator was never below 0.42 times its theoretical centre, so truncating it at one tenth of the centre would leave every realised replication unchanged.

Implementation. All estimators are implemented in R 4.5.1. Standard errors are clustered by city or by province, and tests use t critical values with the number of clusters minus one degrees of freedom. The interaction model is reported only when the target contrast is estimable, which we check from the design matrix before any outcome is simulated. Package versions and settings are recorded in the replication package.

Appendix C. Additional Results

Table A4. Relative bias in the reference design: medians and interquartile ranges across focal programmes.

Estimator	Main: Median	Main: IQR	Literature: Median	Literature: IQR	Half-Synthetic: Median	Half-Synthetic: IQR
TWFE	120%	101% to 161%	118%	108% to 158%	113%	101% to 150%
Event study	100%	66% to 111%	97%	69% to 114%	77%	62% to 104%
CS-never	78%	56% to 110%	79%	56% to 116%	73%	50% to 108%
CS-not-yet	78%	52% to 107%	78%	52% to 112%	72%	47% to 104%
SA	78%	56% to 110%	79%	56% to 116%	73%	50% to 108%
BJS	139%	107% to 175%	138%	114% to 174%	130%	107% to 165%
TWFE + dummies	−54%	−69% to −19%	−55%	−63% to −22%	−51%	−60% to −19%
TWFE + dynamic dummies	−62%	−77% to −45%	−66%	−70% to −44%	−61%	−66% to −50%
TWFE + cohort dummies	−50%	−60% to −20%	−48%	−63% to −21%	−44%	−58% to −20%
Drop-overlapped treated	−119%	−129% to −108%	−104%	−117% to −90%	−109%	−116% to −100%
Stacked, clean controls	89%	76% to 96%	80%	73% to 92%	88%	78% to 97%
Clean both arms	−4%	−12% to 5%	13%	9% to 13%	11%	10% to 17%
Pre-overlap window	−18%	−20% to −13%	−16%	−16% to −4%	−14%	−17% to 2%
Interaction model	−26%	−31% to −11%	−30%	−46% to −13%	−20%	−39% to −2%

Notes: Deletion designs are summarised over the programmes for which they can be estimated: four for drop-overlapped treated and clean both arms on the main and literature scales (five in the half-synthetic design), and five for the pre-overlap window. The half-synthetic columns use the literature-scale reference design.

Table A5. Relative bias/false-positive rate by focal programme, reference design.

Estimator	Scale	Low-Carbon	E-Commerce	New-Energy Vehicles	Sponge	Civilised	Cross-Border
TWFE	Main	98%/11.5%	171%/20.3%	131%/18.3%	7%/4.9%	109%/18.1%	197%/37.7%
TWFE	Literature	108%/95.7%	168%/100.0%	126%/99.4%	17%/2.1%	110%/100.0%	196%/100.0%
CS-not-yet	Main	65%/8.9%	142%/12.6%	90%/29.9%	−3%/3.3%	113%/18.7%	48%/9.7%
CS-not-yet	Literature	70%/63.9%	140%/100.0%	87%/78.6%	−19%/4.0%	121%/100.0%	46%/60.3%
BJS	Main	99%/13.1%	186%/24.6%	133%/18.7%	4%/5.4%	144%/25.9%	206%/40.1%
BJS	Literature	110%/96.8%	183%/100.0%	128%/99.5%	15%/2.3%	147%/100.0%	205%/100.0%
TWFE + dummies	Main	−71%/7.1%	−62%/7.0%	−47%/10.5%	−176%/13.1%	−10%/6.1%	3%/8.4%
TWFE + dummies	Literature	−60%/69.2%	−64%/64.1%	−50%/51.3%	−167%/98.8%	−12%/9.6%	7%/11.5%
Clean both arms	Main	−11%/11.6%	n.a.	2%/13.9%	n.a.	12%/5.4%	−16%/12.6%
Clean both arms	Literature	13%/13.3%	n.a.	−1%/14.2%	n.a.	14%/6.0%	13%/11.4%
Pre-overlap window	Main	−40%/7.8%	−9%/11.8%	−13%/12.8%	n.a.	−20%/6.3%	−18%/13.5%
Pre-overlap window	Literature	−16%/7.0%	−4%/10.2%	−16%/12.8%	n.a.	−19%/6.3%	12%/11.8%
Interaction model	Main	−26%/8.1%	−27%/6.4%	−33%/9.9%	−53%/16.3%	−6%/7.4%	2%/9.5%
Interaction model	Literature	−3%/7.8%	−40%/8.4%	−48%/20.8%	−49%/18.4%	−11%/7.3%	−20%/8.8%

Notes: n.a.: fewer than five treated cities remain.

Table A6. Decomposition of the bias into own-effect weighting and contamination (main scale, medians, % of the focal effect).

Estimator	Total Relative Bias	Own-Effect Weighting	Contamination	Residual
TWFE	120%	0%	118%	2%
Event study	100%	0%	97%	3%
CS-never	78%	0%	78%	0%
CS-not-yet	78%	0%	77%	0%
SA	78%	0%	78%	0%
BJS	139%	0%	137%	1%
TWFE + dummies	−54%	−2%	−55%	3%
TWFE + dynamic dummies	−62%	−3%	−65%	5%
TWFE + cohort dummies	−50%	−3%	−47%	1%
Drop-overlapped treated	−119%	9%	−110%	−17%
Stacked, clean controls	89%	−20%	100%	8%
Clean both arms	−4%	8%	0%	−13%
Pre-overlap window	−18%	−16%	0%	−2%
Interaction model	−26%	−5%	−22%	1%
dCDH	4%	−28%	29%	3%
Stable-status comparisons	−24%	−17%	−2%	−6%

Notes: Main scale, reference design; medians across the focal programmes for which the method can be computed (four for drop-overlapped treated and clean both arms, five for the pre-overlap window, six otherwise). Own-effect weighting is the estimator applied to the noise-free focal signal, minus the ATT; for deletion and stacked designs it includes the change in estimand caused by the retained sample. Contamination is the estimator applied to the noise-free signals of the co-located programmes, divided by the ATT; it is exactly zero for designs whose contamination weights are zero. The residual is the total minus these two terms; it contains Monte Carlo error and, because each column is a separate median, non-additivity across programmes. The dCDH and stable-status rows use the reference-design runs of the main grid; Table A7 reports the additional dCDH runs. Relative contamination does not depend on the common scale of the effects, and for the full-sample methods, the literature-scale totals in Table 2 are within a few percentage points of the main-scale totals.

Table A7. Additional estimators, main-scale reference design.

Estimator	Scale	Pro-grammes	Relative Bias, Full ATT	Relative Bias, Own Estimand	Coverage	False-Positive Rate	Switching Cities Retained
dCDH [31]	Main	6	5%	36%	92%	6.5%	100%
dCDH	Literature	2	18%	66%	90%	44.8%	100%
dCDH, several treatments [12]	Main	6	17%	29%	91%	10.1%	14%
dCDH, several treatments [12]	Literature	2	−13%	4%	84%	17.2%	6.5%
ETWFE [35]	Main	6	78%	–	85%	–	–
dCDH with controls for other treatments	Main	1	−29%	–	97%	3.7%	–
Stable-status comparisons	Main	6	−24%	–	84%	–	–
Control for neighbour exposure	Main	6	−58%	–	93%	7.0%	–
Exclude comparison cities with treated neighbours	Main	6	−23%	–	92%	7.4%	–

Notes: Medians across programmes, city-clustered standard errors. The four dCDH rows are exploratory and come from additional runs with 200 replications for a non-zero and 400 for a zero focal effect (1000 for the standard estimator's main-scale zero-effect design). Their literature-scale rows cover low-carbon and e-commerce cities with co-located effects one and two times the focal effect ($c = 0.4$). "Own estimand" divides the bias by the estimator's response to the noise-free focal signal. "Switching cities retained" is the share of first-period switchers with a valid comparison. With province clustering, the several-treatments false-positive rates are 6.9% (main) and 12.4% (literature). –: not run or not applicable.

The dynamic estimator of de Chaisemartin and D'Haultfœuille [31] appears nearly unbiased, but for offsetting reasons. Its native estimand averages effects over a shorter post-designation window and lies 28% below the full ATT, while contamination adds 29% (Table A6). Additional zero-effect runs put its false-positive rate at 6.5% on the main scale and 44.8% on the literature scale. ETWFE reproduces the point estimates of CS-never exactly in our designs. The two spatial adjustments target spillovers rather than overlap, and both include dummies for the other programmes. Controlling for neighbour exposure therefore behaves like TWFE with dummies, and excluding comparison cities with treated neighbours lowers the bias to −23% without setting the contamination weights to zero.

Table A8. Relative bias in selected main-grid designs (medians across six focal programmes).

Design	TWFE	CS-Not-Yet	BJS	TWFE + Dummies	Clean Both Arms	Pre-Overlap Window	Interaction Model
Homogeneous, $r = 0.5$	92%	54%	96%	−4%	−2%	−6%	−8%
Homogeneous, $r = 1$	169%	103%	177%	2%	−10%	1%	4%
Homogeneous, $r = 2$	331%	215%	358%	0%	6%	−3%	−2%
Cohort-specific, $c = 0$	170%	109%	179%	6%	13%	4%	9%
Cohort-specific, $c = 0.4$	112%	52%	121%	−59%	21%	8%	38%
Cohort-specific, $c = 0.8$	33%	−5%	30%	−146%	33%	1%	63%
Dynamic, $c = 0$	210%	144%	214%	17%	17%	−13%	−42%
Dynamic, $c = 0.4$ (reference)	120%	78%	139%	−54%	−4%	−18%	−26%
Dynamic, $c = 0.8$	41%	14%	41%	−141%	22%	−4%	9%
Dynamic, $c = 0.4$, $s = 0$	95%	45%	112%	−66%	5%	−7%	8%
Dynamic, $c = 0.4$, $s = 0.3$	122%	91%	142%	−54%	8%	−18%	−46%
Reference, $r = 0.5$	58%	37%	67%	−25%	4%	−15%	−17%
Reference, $r = 2$	241%	156%	280%	−109%	−13%	−18%	−60%
Interaction, -0.1σ	−200%	−133%	−207%	−6%	−310%	−242%	−213%
Interaction, $+0.1\sigma$	46%	30%	58%	−41%	−51%	−63%	−75%
Spatial spillovers	126%	85%	147%	−53%	5%	−2%	−30%
One-year anticipation	109%	17%	127%	−64%	13%	−19%	−29%
Selection on trends	216%	111%	253%	−17%	67%	18%	15%
Autocorrelation, 0	122%	65%	143%	−57%	11%	−11%	−39%
Autocorrelation, 0.8	124%	79%	141%	−49%	24%	−8%	−22%

Notes: In the homogeneous and cohort-specific designs, all programmes, the focal one included, have that type of effect. r is the ratio of each co-located programme's average effect to the focal effect; c is the correlation of co-located effects with the focal adoption cohort; s is the dynamic step of co-located effects. Interaction designs measure bias against the observed-state ATT, which includes the interaction terms.

Table A9. Contamination under alternative assumptions about the co-located effects (literature scale, reference design, medians across six focal programmes, % of the focal effect).

Estimator	All Nine	Top Two	Top Three	Random Sizes (5th–95th pct.)	Opposite Sign	Mixed Signs (5th–95th pct.)	FPR, Top Two	FPR, Top Three
TWFE	118	36	58	94 to 184	−118	−46 to 45	26.4	57.2
Event study	97	18	36	77 to 151	−97	−36 to 35	11.9	32.0
CS-never	78	19	34	61 to 123	−78	−36 to 37	9.9	22.2
CS-not-yet	77	18	33	60 to 122	−77	−37 to 36	9.8	21.9
SA	78	19	34	61 to 123	−78	−36 to 35	10.1	22.8
BJS	137	40	66	111 to 213	−137	−49 to 50	30.4	60.6
TWFE + dummies	−55	−23	−32	−88 to −41	55	−22 to 23	15.0	25.4

Table A9. *Cont.*

Estimator	All Nine	Top Two	Top Three	Random Sizes (5th–95th pct.)	Opposite Sign	Mixed Signs (5th–95th pct.)	FPR, Top Two	FPR, Top Three
TWFE + dynamic dummies	−65	−27	−38	−102 to −49	65	−26 to 25	18.7	32.4
TWFE + cohort dummies	−47	−19	−28	−75 to −36	47	−20 to 20	12.5	20.6
Interaction model	−22	−7	−12	−39 to −13	22	−16 to 16	5.4	5.8

Notes: Contamination is computed from each estimator's weights in the reference design and requires no new simulation. All nine: every co-located programme has the focal effect (as in Table A6). Top two, top three: only the two or three programmes with the largest overlap with the focal programme's treated city-years have that effect. Random sizes: each programme's effect is drawn independently as 0.5, 1 or 2 times the focal effect (1000 draws). Opposite sign: every co-located programme has the focal effect with the opposite sign. Mixed signs: four or five of the nine programmes have a positive and the others a negative effect of the focal size (1000 random assignments). Percentiles are taken over draws for each focal programme and then summarised by the median across focal programmes. FPR: false-positive rate implied by Equation (4) with the mean city-clustered standard error of the literature-scale reference design, in percent. Clean both arms and the pre-overlap window have zero contamination in every scenario. Re-simulating 16 estimator–scenario combinations (low-carbon and e-commerce; TWFE, CS-never, TWFE + dummies and the interaction model; top two and opposite sign) with 200 replications each reproduced the predicted biases to within 0.0014 log points.

Table A10. Treated and comparison cities retained by each design.

Design	Low-Carbon	E-Commerce	New-Energy Vehicles	Sponge	Civilised	Cross-Border
A. Full treatment matrix: 297 cities, 2000–2022						
Full sample	70/227	68/229	85/212	27/270	148/149	152/145
Drop-overlapped treated	5/227	2/229 †	6/212	2/270 †	15/149	16/145
Stacked, clean controls	70/185	67/162	85/108	27/91	148/286	151/88
Clean both arms	5/59	2/59 †	6/59	2/59 †	15/59	16/59
Pre-overlap window	26/271	11/286	39/258	4/293 †	60/237	17/280
B. PM2.5 panel: 289 cities, 2000–2021						
Full sample	68/221	65/224	82/207	27/262	144/145	95/194
Drop-overlapped treated	6/221	5/224	8/207	2/262 †	20/145	7/194
Stacked, clean controls	68/181	64/158	82/105	27/88	144/278	94/85
Clean both arms	6/66	5/66	8/66	2/66 †	20/66	7/66
Pre-overlap window	26/263	11/278	38/251	4/285 †	58/231	7/282

Notes: Entries are treated/comparison cities. Treated: cities designated for the focal programme within the retained sample. Comparison: retained cities not designated for the focal programme within the retained sample; in the full sample and the deletion designs, not-yet-designated cities also serve as comparisons before their designation. Drop-overlapped treated removes focal-designated cities that host another programme while designated; clean both arms keeps only cities that never host another programme, so its comparison group consists of the cities without any programme; the pre-overlap window keeps each city's years before its first other programme. For stacked, clean controls, entries count treated cities in supported stacks (three years before and after designation, at least five treated cities) and the distinct clean not-yet- or never-designated cities used as comparisons. † Fewer than five treated cities: the design is not estimated. Counts use the treatment matrix alone. Panel A underlies the simulations (Tables 2, 3, 5 and A4–A9); Panel B the half-synthetic design and Table 4.

References

1. Heilmann, S. Policy experimentation in China's economic rise. *Stud. Comp. Int. Dev.* **2008**, *43*, 1–26. [CrossRef]

2. Wang, S.; Yang, D.Y. Policy experimentation in China: The political economy of policy learning. *J. Political Econ.* **2025**, *133*, 2180–2228. [CrossRef]
3. Ma, X.; Sun, T. Does China's low-carbon city pilot policy effectively enhance urban ecological efficiency? *Sustainability* **2025**, *17*, 368. [CrossRef]
4. Neeley, G.W.; Richardson, L.E. Marijuana policy bundles in the American states over time and their impact on the use of marijuana and other drugs. *Eval. Rev.* **2022**, *46*, 165–199. [CrossRef]
5. Griffin, B.A.; Schuler, M.S.; Pane, J.; Patrick, S.W.; Smart, R.; Stein, B.D.; Grimm, G.; Stuart, E.A. Methodological considerations for estimating policy effects in the context of co-occurring policies. *Health Serv. Outcomes Res. Methodol.* **2023**, *23*, 149–165. [CrossRef]
6. Ann, J.; Lee, M.-J.; Chung, H.J. Finding effects of sequential housing price control policies using various forms of difference in differences. *Eval. Rev.* **2026**, *50*, 482–515. [CrossRef]
7. Goodman-Bacon, A. Difference-in-differences with variation in treatment timing. *J. Econom.* **2021**, *225*, 254–277. [CrossRef]
8. de Chaisemartin, C.; D'Haultfœuille, X. Two-way fixed effects estimators with heterogeneous treatment effects. *Am. Econ. Rev.* **2020**, *110*, 2964–2996. [CrossRef]
9. Callaway, B.; Sant'Anna, P.H.C. Difference-in-differences with multiple time periods. *J. Econom.* **2021**, *225*, 200–230. [CrossRef]
10. Sun, L.; Abraham, S. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *J. Econom.* **2021**, *225*, 175–199. [CrossRef]
11. Borusyak, K.; Jaravel, X.; Spiess, J. Revisiting event-study designs: Robust and efficient estimation. *Rev. Econ. Stud.* **2024**, *91*, 3253–3285. [CrossRef]
12. de Chaisemartin, C.; D'Haultfœuille, X. Two-way fixed effects and differences-in-differences estimators with several treatments. *J. Econom.* **2023**, *236*, 105480. [CrossRef]
13. Goldsmith-Pinkham, P.; Hull, P.; Kolesár, M. Contamination bias in linear regressions. *Am. Econ. Rev.* **2024**, *114*, 4015–4051. [CrossRef]
14. Roth, J.; Sant'Anna, P.H.C.; Bilinski, A.; Poe, J. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *J. Econom.* **2023**, *235*, 2218–2244. [CrossRef]
15. Yang, T.; Tol, R.S.J. Interactions between multiple environmental markets: Addressing contamination bias in overlapping policies. *Energy Econ.* **2026**, *153*, 109094. [CrossRef]
16. Wen, J.; Su, S. An empirical analysis of the synergistic effect of urban pilot policies in China. *Sustainability* **2023**, *15*, 6313. [CrossRef]
17. Yang, L.; Lin, Y. Does dual-pilot policy of Broadband China and low-carbon city enhance carbon emission efficiency? *Sustainability* **2025**, *17*, 10451. [CrossRef]
18. Hedberg, E.C.; Hedges, L.V. Computing statistical power for the difference in differences design. *Eval. Rev.* **2026**, *50*, 149–180. [CrossRef]
19. Miratrix, L.W. Using simulation to analyze interrupted time series designs. *Eval. Rev.* **2022**, *46*, 750–778. [CrossRef]
20. Dee, T.S. The case for preregistering quasi-experimental program and policy evaluations. *Eval. Rev.* **2025**, *49*, 931–945. [CrossRef]
21. Wang, L.; Xie, Q.; Xue, F.; Li, Z. Does smart city construction reduce haze pollution? *Int. J. Environ. Res. Public Health* **2022**, *19*, 16421. [CrossRef]
22. Zou, W.; Pan, M. Does the construction of network infrastructure reduce environmental pollution? Evidence from a quasi-natural experiment in "Broadband China". *Environ. Sci. Pollut. Res.* **2023**, *30*, 242–258. [CrossRef]
23. Zou, X.; Min, J.; Meng, S.; Dai, W.; Li, X. Can digital governance stimulate green development? Evidence from the program of "pilot cities regarding information for public well-being" in China. *Front. Public Health* **2025**, *12*, 1463532. [CrossRef]
24. Li, T.; Wang, L. From security to spending: How China's long-term care insurance pilot drives urban household consumption. *Front. Public Health* **2025**, *13*, 1667168. [CrossRef]
25. Hong, W.; Liu, L. The impact of winter clean heating in northern China on air quality and its spatial spillover effect of the pilot cities. *Sci. Rep.* **2025**, *15*, 36773. [CrossRef]
26. Wang, H.; Shi, X. Low-carbon pilot policy and urban carbon metabolism efficiency: Evidence from China. *Carbon Balance Manag.* **2026**, *21*, 66. [CrossRef]
27. Zhang, L.; Long, T.; Yu, N. The green code of heating revolution: Exploring the impact of China's northern clean heating pilot on the green total factor energy efficiency. *Front. Public Health* **2026**, *14*, 1854174. [CrossRef]
28. Li, B.; Yang, J.; Sun, W. Can expanding cultural consumption improve urban air quality? An analysis based on China's cultural consumption pilot policy. *Int. J. Environ. Res. Public Health* **2023**, *20*, 642. [CrossRef]
29. Zou, S.; Ji, Y. Can digital governance promote urban energy conservation and emission reduction? A quasi-natural experiment based on "National Pilot Policy of Information Benefiting the People" in China. *PLoS ONE* **2025**, *20*, e0320007. [CrossRef]
30. Deng, C.; Guo, B. The effects and mechanisms of low-carbon transition promoting air quality improvement: Based on empirical evidence of low-carbon city pilot policy. *Front. Public Health* **2026**, *14*, 1799779. [CrossRef]
31. de Chaisemartin, C.; D'Haultfœuille, X. Difference-in-differences estimators of intertemporal treatment effects. *Rev. Econ. Stat.* **2026**, *108*, 863–880. [CrossRef]

32. Wei, J.; Li, Z.; Cribb, M.; Huang, W.; Xue, W.; Sun, L.; Guo, J.; Peng, Y.; Li, J.; Lyapustin, A.; et al. Improved 1 km resolution PM2.5 estimates across China using enhanced space–time extremely randomized trees. *Atmos. Chem. Phys.* **2020**, *20*, 3273–3289. [CrossRef]
33. Wei, J.; Li, Z.; Lyapustin, A.; Sun, L.; Peng, Y.; Xue, W.; Su, T.; Cribb, M. Reconstructing 1-km-resolution high-quality PM2.5 data records from 2000 to 2018 in China: Spatiotemporal variations and policy implications. *Remote Sens. Environ.* **2021**, *252*, 112136. [CrossRef]
34. Yan, J.; Zhao, J.; Yang, X.; Su, X.; Wang, H.; Ran, Q.; Shen, J. Does low-carbon city pilot policy alleviate urban haze pollution? Empirical evidence from a quasi-natural experiment in China. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11287. [CrossRef]
35. Wooldridge, J.M. Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. *Empir. Econ.* **2025**, *69*, 2545–2587. [CrossRef]
36. Roth, J. Pretest with caution: Event-study estimates after testing for parallel trends. *Am. Econ. Rev. Insights* **2022**, *4*, 305–322. [CrossRef]
37. Arkhangelsky, D.; Athey, S.; Hirshberg, D.A.; Imbens, G.W.; Wager, S. Synthetic difference-in-differences. *Am. Econ. Rev.* **2021**, *111*, 4088–4118. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.