A Response to Green and Aronow (2026)*

Yiqing Xu

April 26, 2026

Abstract

Green and Aronow (2026), hereafter G&A, document that the `gsynth` R package, prior to v1.3.1 (December 2025), applied an undocumented parametric-bootstrap routine under the IFE-EM estimator of Gobillon and Magnac (2016) that omitted the out-of-sample leave-one-out step Xu (2017) prescribed for the generalized synthetic control (GSC) estimator. Under serially correlated errors the consequence is drastic undercoverage. I acknowledge the error and agree with the substance of G&A's analysis. This note develops three further points. First, the leave-one-out fix alone does not rescue the parametric bootstrap procedure for IFE-EM: its factor estimate depends on treated-unit data, so the prediction errors are not truly out-of-sample and the bootstrap remains miscalibrated. Second, serially correlated data drive cross-validation (CV) to select too many factors, and this overfitting compounds with the bootstrap implementation error: the regimes that drive CV to over-select are exactly those in which the bootstrap most under-covers. Third, I reanalyze the three APSR applications and broadly concur with G&A's substantive conclusions, while drawing further lessons for applied work on visual inspection and simple diagnostics. New safeguards are introduced in the updated `fect` package.

*Yiqing Xu, Assistant Professor in Political Science, Stanford University. Email: yiqingxu@stanford.edu. I thank Beniamino Green and P. M. Aronow for identifying the implementation error in the `gsynth` R package documented below, and for the care with which they documented its consequences.

1. Introduction and Acknowledgment

The parametric-bootstrap procedure of Xu (2017) was proposed for settings with a small number of treated units (N_{tr})—the regime in which synthetic control methods are most often applied—where the nonparametric bootstrap that resamples all units is infeasible. It has since been routinely used in applied panel work. G&A document that pre-v1.3.1 `gsynth` applied an undocumented variant of this procedure when the IFE-EM estimator (Gobillon and Magnac, 2016) was selected: bootstrap samples were constructed from in-sample residuals only, without the out-of-sample leave-one-out loop—Step 1 of Xu (2017)'s Algorithm 2—that Xu (2017) prescribed for GSC. Xu (2017) did not specify an inferential procedure for IFE-EM, and the `gsynth` implementation for the parametric bootstrap procedure for IFE-EM was never separately documented. I acknowledge this implementation error and regret it. It was present in the `gsynth` code base from its first CRAN release in March 2017 through v1.3.0, and was removed during the December 2025 refactor that absorbed `gsynth` into `fect` v2.0.0 and issued `gsynth` v1.3.1.

Under serially correlated errors and model overfitting, the consequence of the error is not a small numerical discrepancy but a sharp collapse in the bootstrap variance. G&A's "Toy Example" (pp. 7–9) makes the mechanism vivid: on a zero-effect data-generating process with Gaussian-kernel error correlation, cross-validation (CV) selects seven factors despite there being zero true factors; the fit attains an in-sample R^2 of 0.997; the residuals fed into the bootstrap vanish; and the confidence interval shrinks to near-zero width. Their empirical Monte Carlo on 16 state-panel outcomes yields average coverage of 49% at nominal 95%, and their reanalysis of three *American Political Science Review* papers—Gilens et al. (2021), Alsaadi (2025), and Eibl and Hertog (2023)—shows that most of the originally significant findings do not survive either IFE-EM with the corrected inferential procedure or GSC. Their diagnosis is correct; their evidence is compelling; and their reanalysis is a service to the applied community.

The remainder of this note does not relitigate what G&A have established. Instead, it develops the three points laid out in the abstract. Section 2 fixes notation, distinguishes the two estimators implemented in `gsynth`, lists the points of agreement with G&A, and isolates the narrow question they leave open. Section 3 takes up that question and addresses a second, independent issue. It shows that: (a) under IFE-EM, the leave-one-out correction is insufficient for two structural reasons, while the same procedure works for GSC (§3.1–§3.2), and explains why the error persisted for years (§3.3); and (b) block CV over-selects rank under serial correlation (§3.4). Together, these problems can produce the over-confident intervals that applied work has carried. Section 4 reanalyzes the three APSR papers using `gsynth`, with diagnostics that surface these failures directly from the panel data. Section 5 provides practical recommendations and discusses the broader lessons from this exchange.

2. Setup and Notation

I follow the notation of Xu (2017). Let $i \in \{1, \ldots, N\}$ index units and $t \in \{1, \ldots, T\}$ index time periods, and let $D_{it} \in \{0, 1\}$ denote treatment status. Let $\mathcal{T} = \{i : D_{it} = 1 \text{ for some } t\}$ be the set of treated units and $\mathcal{C} = \{1, \ldots, N\} \setminus \mathcal{T}$ the set of control units, with cardinalities $N_{\text{tr}} = |\mathcal{T}|$ and $N_{\text{co}} = |\mathcal{C}|$. For exposition I take the block-treatment case in which all units in $\mathcal{T}$ adopt treatment at the same date T_0; the analysis below generalizes to the staggered case without structural change. Let $Y_{it}(0)$ and $Y_{it}(1)$ denote potential outcomes, with observed outcome $Y_{it} = D_{it}Y_{it}(1) + (1 - D_{it})Y_{it}(0)$, and let $\text{ATT}_t = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid i \in \mathcal{T}, D_{it} = 1]$ for $t > T_0$. The factor model for the untreated outcome is

$$Y_{it}(0) = x_{it}'\beta + \lambda_i' f_t + \varepsilon_{it}, \tag{1}$$

where $\lambda_i \in \mathbb{R}^r$ is a unit-specific loading, $f_t \in \mathbb{R}^r$ is a time-specific factor, x_{it} is a vector of covariates with parameter β, and ε_{it} is an idiosyncratic error.

The two estimators available in `gsynth` differ in exactly one respect: whether treated-unit

pre-treatment cells enter factor estimation. The GSC estimator of Xu (2017) solves

$$(\hat{F}, \hat{\Lambda}_{\mathcal{C}}, \hat{\beta}) \;=\; \arg\min_{F,\Lambda,\beta} \sum_{i\in\mathcal{C}} \sum_{t=1}^{T} (Y_{it} - x_{it}'\beta - \lambda_i' f_t)^2 \tag{2}$$

over the control sample only, subject to standard normalizations. Treated-unit loadings are then estimated by projection onto $\hat{F}$ using pre-treatment data only,

$$\hat{\lambda}_i \;=\; (\hat{F}_{\text{pre}}'\hat{F}_{\text{pre}})^{-1}\hat{F}_{\text{pre}}'(Y_i^{\text{pre}} - X_i^{\text{pre}}\hat{\beta}), \quad i \in \mathcal{T}. \tag{3}$$

The IFE-EM estimator of Gobillon and Magnac (2016) estimates factors and all loadings jointly over both control cells and treated pre-treatment cells,

$$(\hat{F}, \hat{\Lambda}, \hat{\beta}) \;=\; \arg\min_{F,\Lambda,\beta} \sum_{(i,t):\, O_{it}=1} (Y_{it} - x_{it}'\beta - \lambda_i' f_t)^2, \tag{4}$$

where $O_{it} = 1 - D_{it}$ is the observation indicator for the untreated outcome (so $O_{it} = 1$ on all control cells and all treated-unit pre-treatment cells). Treated post-period cells are iteratively imputed via expectation-maximization (hence the name, IFE-EM).

The structural difference between Equations (2) and (4) is whether the treated-pre block $\{(i,t) : i \in \mathcal{T}, t \le T_0\}$ appears in the factor-estimation objective. Under GSC it does not; under IFE-EM it does. Figure 1 illustrates. This single structural difference is what drives the analysis of Section 3.

Asymptotic unbiasedness. When the factor model in Equation (1) is correctly specified—including the number of factors r—both GSC and IFE-EM are asymptotically unbiased for ATT_t as $T_0 \to \infty$ and $N \to \infty$. This differs from the conventional consistency notion: because N_{tr} is small and held fixed, the average idiosyncratic shock across treated units does not vanish in the limit, so the familiar $\sqrt{N}$-scaling does not apply. Uncertainty quantification in this regime has to be done differently from a standard M-estimation setting, and this is the

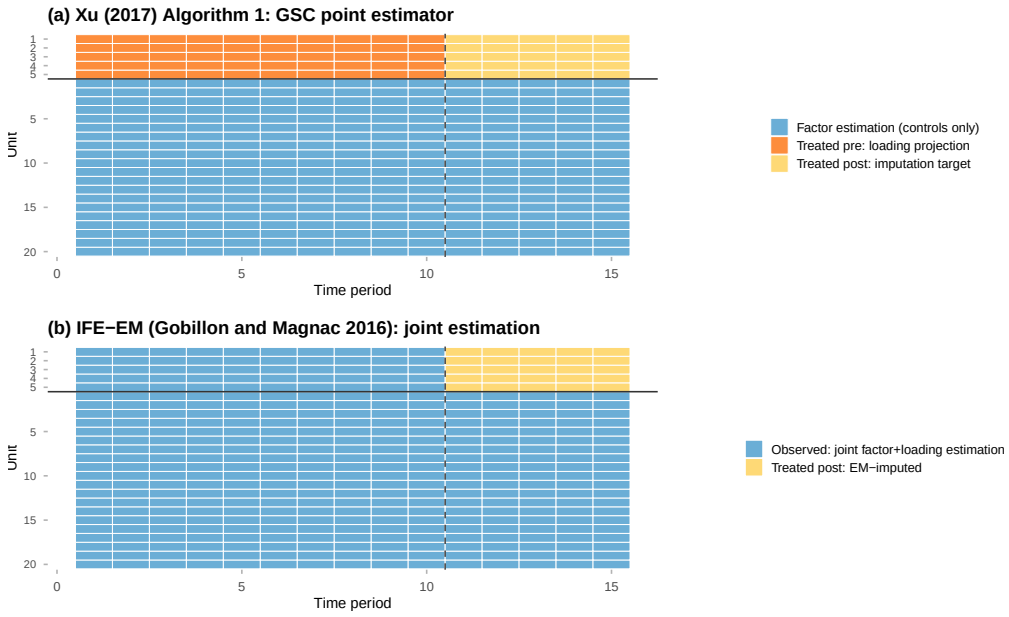


Figure 1: Data usage for the two factor-model estimators implemented in `gsynth`. Treated units are at the top of each panel, controls below; the dashed vertical line marks treatment onset at $t = T_0 + 0.5$. **(a)** GSC: factors $\hat{F}$ are estimated from control-unit cells only (blue); treated-unit loadings are then estimated by projection onto $\hat{F}$ using treated pre-treatment cells (orange); treated post-period cells are the target of counterfactual imputation (yellow). **(b)** IFE-EM: factors and loadings are estimated jointly over all observed cells, including treated-unit pre-periods (blue on both); treated post-period cells are iteratively imputed via EM (yellow). The only structural difference is in the top-left block: orange in (a), blue in (b).

reason the Xu (2017) parametric bootstrap was proposed for GSC: it targets the conditional variance V_t defined below, not the sampling variance of a consistent estimator.

Target of inference. The target is the sampling distribution of the ATT estimator conditional on the factors, loadings, covariates, and treatment assignment. Fixing $t > T_0$,

$$V_t \;=\; \mathrm{Var}\Big(\widehat{\mathrm{ATT}}_t - \mathrm{ATT}_t \;\Big|\; \Lambda, F, X, D\Big)\,, \tag{5}$$

where the variance integrates only over the idiosyncratic errors $\{\varepsilon_{it}\}$. This is the target Xu (2017) states.

Consider California's Proposition 99 as a concrete example: California is the single treated unit and the 38 non-Prop-99 states are the controls. Conditioning on (Λ, F, X, D) treats as given California, the control set, the covariates, the latent factor trajectory over 1970–

2000, California's loading on those factors, and the 1988 treatment date; only the state-year idiosyncratic shock ε_{it} is integrated over. V_t answers "how noisy is the post-1988 gap estimate for *this* panel?" not "what if Prop 99 had passed in a different state?" With $N_{\text{tr}} = 1$, there is no population of hypothetical Californias to average over, hence, this conditional variance is also a plausible way to move forward. A complementary direction in ongoing work is conformal inference on the counterfactual prediction error, which yields valid prediction intervals for the post-treatment gap for finite samples.

2.1. Why the implementation error causes undercoverage

Decomposing the estimation error,

$$\widehat{\text{ATT}}_t - \text{ATT}_t = \underbrace{\frac{1}{N_{\text{tr}}} \sum_{i \in \mathcal{T}} (\lambda_i - \hat{\lambda}_i)' \hat{f}_t}_{\text{loading error}} + \underbrace{\overline{\lambda}_{\mathcal{T}}' (f_t - \hat{f}_t)}_{\text{factor error}} + \underbrace{x_t^\top (\beta - \hat{\beta})}_{\text{covariate error}} + \bar{\varepsilon}_t^{\text{tr}}, \tag{6}$$

where $\bar{\varepsilon}_t^{\text{tr}} = N_{\text{tr}}^{-1} \sum_{i \in \mathcal{T}} \varepsilon_{it}$. A valid bootstrap must recover the joint distribution of all four terms. In-sample residuals on control cells reflect only the fourth; the leave-one-out prediction error $\hat{\varepsilon}_{i^*}^p = Y_{i^*} - \hat{Y}_{i^*}^{(-i^*)}(0)$ is richer, combining noise, nuisance-estimation error in $(\hat{F}, \hat{\Lambda}, \hat{\beta})$, and any residual misspecification of the factor model.

The reason this richer pool delivers honest coverage even when the factor rank is over-selected is that the held-out unit i^* contributes nothing to the fit. An over-parametrized model drives in-sample residuals toward zero by absorbing noise patterns as factor structure, so the empirical variance of in-sample residuals understates $\text{Var}(\varepsilon_{i,t})$. The spurious factors that fit training-unit noise, however, are uninformative about i^*, whose data they have not seen; the prediction error $\hat{\varepsilon}_{i^*}^p$ remains genuine and reflects, if anything, additional projection-error variance from the over-parametrized fit. Resampling these prediction errors therefore yields a bootstrap variance that does not collapse under rank over-selection. This is what Step 1 of Xu (2017)'s Algorithm 2 is designed to supply for GSC. G&A reproduce Algorithm 2 on

their page 5 and the pre-v1.3.1 IFE-EM bootstrap on their page 7; Appendix A.1 records all five bootstrap procedures of this note as pseudocode.

The pre-v1.3.1 `gsynth` bootstrap for IFE-EM omitted Step 1 and drew its perturbations from in-sample residuals alone. A residual bootstrap uses the empirical residual variance $\hat{\sigma}^{*2}$ in place of σ^2 throughout, so every contribution to V_t that scales with σ^2 in truth scales with $\hat{\sigma}^{*2}$ in the bootstrap: the treated-noise term contributes σ^2/N_{tr} to V_t and $\hat{\sigma}^{*2}/N_{\text{tr}}$ to its bootstrap analog, and the loading, factor, and covariate-error terms enter through refits on perturbed data $Y^* = \hat{Y} + \varepsilon^*$, where each refit estimate is approximately linear in ε^* and so its across-replication variance also scales with $\hat{\sigma}^{*2}$. Whenever CV selected a rank larger than the noise structure could support—which, as G&A's "Toy Example" shows, happens routinely under serial correlation—$\hat{\sigma}^{*2}$ fell well below σ^2 and the bootstrap CI width contracted by $\hat{\sigma}^*/\sigma$ uniformly. This is the immediate mechanism behind the undercoverage G&A document.

2.2. Points of agreement and the open question

I take the following contributions of G&A as established and do not reproduce them.

- The identification of the implementation error (G&A, Algorithm 2 on p. 6; source-code citations in footnote 2 of their text).
- The "Toy Example" showing how CV under serial correlation selects too many factors and collapses the bootstrap variance (pp. 7–8).
- The empirical evidence of this failure is clear: 49% coverage at nominal 95% across 16 state-panel outcomes (pp. 8–9), and substantive changes in the reanalyses of the three empirical applications to varying degrees. Section 4 reproduces these results using `fect`.

G&A explicitly leave open whether the Xu (2017) parametric bootstrap is appropriate for IFE-EM even once the implementation error is corrected: "we remain uncertain about the conditions under which Xu (2017)'s parametric bootstrap is appropriate for inference in IFE models. We are not aware of theory suggesting that Xu (2017)'s parametric bootstrap would

be suitable when researchers use the GSC method to impute counterfactuals, but unsuitable when the IFE-EM method is used for such imputation" (G&A, p. 6). §3.1–§3.2 address this question; §3.4 takes up the related rank-selection issue.

3. Two Issues with the Residual Bootstrap

G&A propose to restore Step 1 of Xu (2017)'s Algorithm 2: the leave-one-out loop that, under GSC, generates out-of-sample prediction errors from the control sample and feeds them into the residual-bootstrap in Step 2. Under GSC, this procedure works as documented in the original paper. Under IFE-EM, it is not, and serial correlation introduces a separate failure mode the LOO step does not address.

3.1. Why the LOO correction does not rescue IFE-EM

Step 1 of Algorithm 2 holds out one control unit $i^* \in \mathcal{C}$, re-fits the factor model on the remaining sample, and records the held-out unit's prediction error $\hat{\varepsilon}^p_{i^*} = Y_{i^*} - \hat{Y}^{(-i^*)}_{i^*}(0)$. Under GSC, the re-fit uses the controls $\mathcal{C} \setminus \{i^*\}$ alone, so the factor space $\hat{F}^{(-i^*)}$ against which the held-out unit is evaluated is an estimator constructed from data that does not include i^*; the held-out residual is honestly out-of-sample. Under IFE-EM, two distinct mechanisms break this property.

First, treated-pre cells contaminate the leave-one-out factor space. The IFE-EM re-fit solves

$$(\hat{F}^{(-i^*)}, \hat{\Lambda}^{(-i^*)}, \hat{\beta}^{(-i^*)}) \;=\; \arg\min \underbrace{\sum_{i \in \mathcal{C}\setminus\{i^*\}} \sum_{t=1}^{T} (\,\cdot\,)^2}_{\text{controls minus } i^*} \;+\; \underbrace{\sum_{i \in \mathcal{T}} \sum_{t \le T_0} (\,\cdot\,)^2}_{\text{treated-pre block}}, \tag{7}$$

where the IFE-EM objective in Equation (4) retains the treated-pre block regardless of which control is held out. The factor space $\hat{F}^{(-i^*)}$ therefore still depends on the full block $\{(i,t) : i \in \mathcal{T}, t \le T_0\}$; "leaving one out" has been implemented on the control side only. The

held-out residual decomposes as

$$\hat{\varepsilon}_{i^*t}^{p,\mathrm{IFE}} \;=\; \varepsilon_{i^*t} \;+\; (\lambda_{i^*} - \hat{\lambda}_{i^*}^{(-i^*)})'\hat{f}_t^{(-i^*)} \;+\; \overline{\lambda}'(f_t - \hat{f}_t^{(-i^*)}) \;+\; x_{i^*t}^{\top}(\beta - \hat{\beta}^{(-i^*)}),^{1} \tag{8}$$

and $\hat{f}_t^{(-i^*)}$ inherits, through Equation (7), the signal in the treated-pre cells. The bootstrap Loop 2 then refits IFE-EM on pseudo-datasets that again include the same treated-pre cells as inputs to factor estimation; the bootstrap variability over replications fails to integrate over the treated-pre contribution to $\hat{F}$. The "out-of-sample" residual is not out-of-sample in the sense required by the argument in Xu (2017).

Second, the residual pool feeding the leave-one-out step is itself shrunken. When CV, under serial correlation, selects a large rank, the estimated factors absorb the systematic time-variation of the errors, and the in-sample residuals satisfy

$$\mathbb{E}[\hat{\varepsilon}_{it}^2] \;\approx\; \sigma^2\left(1 - \frac{d_{\mathrm{eff}}}{NT}\right), \tag{9}$$

with d_{eff} the trace of the hat matrix scaling in r. The first mechanism above prevents the LOO step from escaping this shrinkage under IFE-EM: $\mathcal{T}_{\mathrm{pre}}$ continues to shape $\hat{F}^{(-i^*)}$, so the prediction errors inherit the in-sample contraction. The EM step additionally inflates d_{eff} beyond the nominal $(N+T)r$ through the imputed treated-post block. The core failure G&A identify—residuals that vanish because the factor basis absorbs the serial correlation—is propagated one level up under IFE-EM, not resolved by the leave-one-out step.

The two mechanisms compound. The first affects the factor space against which the held-out residual is measured; the second affects the residual itself, regardless of whether it is measured in-sample or out-of-sample. Correcting either one without the other leaves the other in place. A leave-one-out correction, applied without modification to the IFE-EM fit, acts only at the level of where residuals are collected, not at the level of how shrunken they

[1]Throughout this response, $\overline{\lambda} \equiv N_{\mathrm{tr}}^{-1}\sum_{i\in\mathcal{T}}\lambda_i$ denotes the average treated loading. Per-unit factor-error decompositions of the form $\lambda_{i^*}'(f_t - \hat{f}_t)$ are replaced by $\overline{\lambda}'(f_t - \hat{f}_t)$ as a leading-order approximation in the small-N_{tr} regime.

are.

Sample splitting. A natural further modification is to split the control pool $\mathcal{C}$ into two halves $\mathcal{C}_A, \mathcal{C}_B$, estimate the factor space from $\mathcal{C}_A$, and collect residuals from $\mathcal{C}_B$. This eliminates overlap between the factor-estimation sample and the residual-collection sample on the control side. It does not, however, sever the treated-pre contamination: the treated-pre block remains in the factor-estimation objective regardless of how $\mathcal{C}$ is partitioned. The split-factor estimator $\hat{F}^{(A)}$ still depends on $\mathcal{T}_{\text{pre}}$, and the residuals collected on $\mathcal{C}_B$ are still evaluated against a factor basis shaped by data that will re-appear in the bootstrap loop. Appendix A.2.3 gives the formal account.

Monte Carlo evidence Table 1 reports empirical coverage of nominal 95% confidence intervals under four data-generating processes.[2] The three DGPs in the upper panel have serially correlated errors: one uses the long-range Gaussian-kernel covariance from G&A's "Toy Example" (no factor structure); the other two use the Xu (2017) $r = 2$ factor structure with AR(1) errors at $\rho = 0.8$, differing only in whether the rank used in fitting equals the true rank or exceeds it. The single DGP in the lower panel has i.i.d. errors at the Xu factor structure with the rank correctly specified. Across all three serial-correlation DGPs, the leave-one-out-corrected Variant (ii) tracks the pre-v1.3.1 Variant (i) to within two or three percentage points of coverage; neither procedure comes close to nominal. Under i.i.d. errors with the rank correctly specified, all three procedures cover at or near nominal. This last row is the reason the implementation error survived years of validation: the standard simulation design in the factor-model literature uses i.i.d. errors, which is the one regime in which the

[2]Each replication draws a fresh $(\Lambda, F, \alpha, \xi)$; only the block-treatment indicator D is held fixed. The law of total variance gives $\text{Var}(\widehat{\text{ATT}}_t - \text{ATT}_t) = \mathbb{E}_{(\Lambda,F)}[V_t(\Lambda, F, X, D)] + \text{Var}_{(\Lambda,F)}[b_t]$, where $b_t = \mathbb{E}_\varepsilon[\widehat{\text{ATT}}_t - \text{ATT}_t \mid \Lambda, F, X, D]$ is the finite-sample conditional bias. Because the second term is weakly positive, the empirical variance across replications upper-bounds $\mathbb{E}_{(\Lambda,F)}[V_t]$: the Monte Carlo is a conservative test of whether the procedure covers V_t on an average realization. Empirical undercoverage therefore implies the procedure under-delivers against V_t end-to-end—a combination of (a) the shortfall is the bootstrap's approximation of V_t and (b) finite-sample conditional bias—and cannot be an artifact of averaging across individually well-calibrated realizations.

pathology does not manifest. GSC with the Xu (2017) parametric bootstrap, finally, is at or above nominal in every cell—an observation that motivates §3.2 below.

DGP description	GSC (Xu Alg. 2)	IFE-EM (i) (pre-v1.3.1 bug)	IFE-EM (ii) (G&A fix)
Serial-correlation regime:			
Long-range ε correlation (G&A toy)	0.960	0.355	0.365
Xu $r = 2$, AR(1) $\rho = 0.8$, rank correct	0.920	0.525	0.550
Xu $r = 2$, AR(1) $\rho = 0.8$, rank over-specified	0.985	0.380	0.405
Independent-errors regime:			
Xu $r = 2$, i.i.d. ε, rank correct	0.950	0.975	0.970

Table 1: Empirical coverage of nominal 95% confidence intervals under four data-generating processes. Variant (i) is the pre-v1.3.1 `gsynth` implementation; Variant (ii) restores the leave-one-out step of Xu (2017)'s Algorithm 2 as G&A propose. Shared settings: $N_{\text{tr}} = 5$, $T = 30$, $T_0 = 20$, $N_{\text{co}} = 50$; 200 Monte Carlo replications per cell; $B = 100$ bootstrap replicates per replication. Under any of the three serial-correlation DGPs, the leave-one-out correction improves coverage by at most two to three percentage points and does not rescue inference under IFE-EM. Under i.i.d. errors with the rank correctly specified, all three procedures cover at or near nominal. Full Monte Carlo evidence, SE ratios, and N_{tr}-scaling experiments are in Appendix A.3.

3.2. Why the story is different under GSC

The validity of the Xu (2017) procedure under GSC rests on one structural property of the GSC objective.

Insulation property. Under Equation (1) with $\varepsilon_{it} \perp\!\!\!\perp (D_{it}, \lambda_i, f_t)$, the GSC estimators $(\hat{F}, \hat{\Lambda}_{\mathcal{C}}, \hat{\beta})$ defined in Equation (2) depend on $\{Y_{it}, x_{it}\}_{i \in \mathcal{C}}$ only—immediate from inspection of the objective—and are therefore independent of the treated sample $\mathcal{T}$ conditional on $\{f_t\}_{t=1}^{T}$. Factor estimation is insulated from treated-unit data.

The consequence is that the leave-one-out residuals collected in Step 1 of Algorithm 2 are honestly out-of-sample: the held-out unit is not used in the factor fit on the remaining controls, and the factor basis against which its residual is evaluated is not influenced by treated-unit data either. Neither of the two mechanisms from §3.1 arises here. Treated-pre cells cannot contaminate $\hat{F}$ because they do not enter the GSC objective. In-sample residuals on control cells can still exhibit the rank-overfit shrinkage of Equation (9) when CV selects a

large r, but Step 1 does not draw from them—it resamples prediction errors on units whose data did not shape $\hat{F}$, and these prediction errors measure honest out-of-sample uncertainty regardless of in-sample overfit. The parametric bootstrap for GSC inherits its robustness from this predictive-based construction, not from the absence of overfitting.

Sample splitting is not needed under GSC (right now). The double-machine-learning literature (Chernozhukov et al., 2018) calls for sample splitting (cross-fitting) when the nuisance is estimated flexibly—typically nonparametrically—because at slow convergence rates the nuisance error would contaminate asymptotic inference on the target parameter if both were computed on the same sample. Here, the factor model is parametric at a fixed rank r: in the asymptotic regime of §2 ($N_{\text{co}} \to \infty$ and $T_0 \to \infty$ with N_{tr} held fixed), $(\hat{F}, \hat{\Lambda}_{\mathcal{C}}, \hat{\beta})$ converge at rate $\min(\sqrt{N_{\text{co}}}, \sqrt{T_0})$ (Bai, 2003)—fast enough that the first-order bias DML cross-fitting is designed to eliminate does not arise. What the parametric bootstrap still requires—honest out-of-sample prediction errors for treated units—is delivered by the leave-one-out step in Algorithm 2 Step 1 together with the insulation property: treated residuals are evaluated against a factor basis that has not seen them.

To confirm empirically, the Monte Carlo in Appendix A.3 includes a GSC+split variant that partitions the control pool, estimates factors on one half, and collects residuals from the other; as expected, it delivers essentially no coverage improvement over GSC without splitting across all seven DGPs (Tables A1 and A2). A relaxation to a more flexible imputation—matrix completion with a data-driven rank penalty, or a nonlinear/ML imputer—would reintroduce the DML concern, and some form of cross-fitting would again become relevant for uncertainty quantification.

The normal approximation is (usually) innocuous. Two places in the parametric-bootstrap procedure invoke a Gaussian device. First, on unbalanced panels a treated unit's bootstrap error is drawn from $\mathcal{N}(0, \widehat{\text{Vcov}}_{\text{tr}})$ rather than resampled empirically. Second, from v1.1.7 of `gsynth`, confidence intervals are constructed by a normal wrapper around

the bootstrap standard error rather than by percentile. Under the insulation property, $\widehat{\text{Vcov}}_{\text{tr}}$ is consistent at rate $O(N_{\text{co}}^{-1/2})$ because the residuals feeding it are honestly out-of-sample. I introduced this device mainly for a practical reason: at small N_{co} the leave-one-out pool of at most N_{co} residual vectors is coarse, and resampling inherits that coarseness, whereas $\mathcal{N}(0, \widehat{\text{Vcov}}_{\text{tr}})$ supplies a continuous distribution matched to the pool's second-moment structure.

Non-normality of ε_{it} only matters at higher order: it distorts a one-sided tail at rate $O(N_{\text{co}}^{-1/2})$, but the skewness contribution to a symmetric two-sided 95% CI is equal in the two tails and cancels, leaving coverage error of $O(N_{\text{co}}^{-1})$. The caveat is that both devices share an assumption the resampled scheme also needs: control residuals are exchangeable with a treated unit's error; unit-level heteroscedasticity would break both. A coverage simulation at $N = 50$, $T = 20$, $r = 2$, $N_{\text{tr}} = 10$, reported in Appendix A.3, yields empirical coverage in $[0.93, 0.96]$ across three GSC-family specifications in `fect`—a spread consistent with the predicted $O(N^{-1})$ rate; larger N produces values closer to 0.95.

3.3. Why the error was not caught

The pre-v1.3.1 parametric bootstrap for IFE-EM was not derived from a bootstrap-consistency result. I implemented it as the natural residual-bootstrap analogue of Algorithm 2 of Xu (2017)—which addresses GSC—and tested it on the standard factor-model validation regime: i.i.d. errors at correctly specified rank. In that regime, Table 2 reports coverage in $[0.92, 0.98]$ at both $N_{\text{co}} = 50$ and $N_{\text{co}} = 100$, with slow degradation under severe rank over-specification. The procedure passed validation.

It passed for the wrong reason. The SE-ratio column shows that at $N_{\text{co}} = 50$ the bootstrap SE is about 1.74 times the empirical SD of $\widehat{\text{ATT}}_t$ across replications for Variants (i) and (ii), against 1.05 for GSC. The IFE-EM CIs are roughly 70% wider than they should be; they cover at the nominal rate because they are too wide, not because they estimate V_t correctly. At $N_{\text{co}} = 100$ the ratio falls to about 1.07.

Rank specification	Metric	GSC (Xu Alg. 2)	IFE-EM (i) (pre-v1.3.1)	IFE-EM (ii) (G&A fix)
Panel (a): $N_{co} = 50$				
$r_{\text{fit}} = 2$ (correct)	coverage	0.950	0.975	0.970
	SE ratio	1.05	1.74	1.75
$r_{\text{fit}} = 4$ (mild over-spec)	coverage	0.900	0.940	0.955
	SE ratio	0.77	1.50	1.53
$r_{\text{fit}} = 6$ (severe over-spec)	coverage	0.905	0.965	0.960
	SE ratio	0.79	1.62	1.70
Panel (b): $N_{co} = 100$				
$r_{\text{fit}} = 2$ (correct)	coverage	0.955	0.945	0.945
	SE ratio	1.03	1.07	1.05
$r_{\text{fit}} = 4$ (mild over-spec)	coverage	0.955	0.935	0.940
	SE ratio	1.01	1.03	1.02
$r_{\text{fit}} = 6$ (severe over-spec)	coverage	0.930	0.860	0.865
	SE ratio	0.81	0.76	0.76

Table 2: Monte Carlo coverage and SE ratio under i.i.d. errors at the Xu (2017) $r = 2$ factor structure. Panel (a) at $N_{\text{co}} = 50$; Panel (b) at $N_{\text{co}} = 100$. The SE ratio is the mean bootstrap SE divided by the empirical SD of $\widehat{\text{ATT}}_t$ across replications; a correctly calibrated procedure has SE ratio ≈ 1. Under i.i.d. errors with correctly or mildly over-specified rank, IFE-EM Variants (i) and (ii) attain near-nominal coverage through over-conservative intervals (SE ratios well above one at $N_{\text{co}} = 50$); the over-width shrinks toward calibration as N_{co} grows. Shared settings: $N_{\text{tr}} = 5$, $T = 30$, $T_0 = 20$; 200 replications per cell; $B = 100$ bootstrap replicates. Full results including GSC with control-pool sample splitting and the exploratory Variant (iii) are in Appendix A.3, Table A2.

The source of the over-width is a structural mismatch between what the bootstrap perturbs and what the original estimator treats as stochastic. Variant (i) adds residual draws ε^* to both control cells and the imputed counterfactuals $\hat{Y}_{it}(0)$ on treated-post cells, then re-fits IFE-EM. But in the original fit, those imputed cells are deterministic given the fitted parameters—at EM convergence the imputed value equals the fitted value, and the residual is zero by construction. The bootstrap therefore introduces variability that was not part of $\widehat{\text{ATT}}_t$'s original sampling distribution; the refit absorbs it into $\hat{F}^{(k)}$, and $\widehat{\text{ATT}}_t^{(k)}$ inherits the excess through its dependence on the factor basis. GSC avoids this because factor estimation uses only control cells in both the original fit and the refit (§3.2). The relative contribution of treated-post perturbations shrinks as N_{co} grows, which is why the SE-ratio contracts toward one.

This sketches a mechanism; it is not a proof. The consistency of the IFE estimator itself under i.i.d. errors with correctly specified rank is given by Bai (2009), which secures the residual pool's convergence to the true error distribution—the first step of any residual-bootstrap argument. The validity of a residual bootstrap for factor-augmented regressions, in the sense of matching the asymptotic distribution of a downstream OLS coefficient, has been formally established by Gonçalves and Perron (2014). Their setting—factors estimated from one panel, used as regressors in a separate observed-outcome regression—is structurally related to but distinct from Variant (i), where factor estimation and ATT estimation share one objective, the treated-post cells of that objective are never observed, and bootstrap perturbations on imputed cells have no counterpart in their framework. A bootstrap-consistency result for the joint-objective imputation setup is not, to my knowledge, in the literature. The near-nominal coverage in Table 2 is consistent with asymptotic validity but does not establish it.

Under serial correlation the over-width inverts: the SE-ratio falls to 0.22–0.38 (Table A1) because the IFE-EM factor space absorbs the systematic time-variation of the errors and the in-sample residuals contract far below the $\sqrt{1 - d/NT}$ their nominal parameter count would predict. The validation I performed did not include this regime, and would have surfaced the problem if it had. The procedure is not reliably calibrated under either regime—the i.i.d. regime hides this through over-conservative intervals, the serial-correlation regime exposes it through undercoverage—which is why the current policy in `fect` disallows the parametric bootstrap under IFE-EM.

3.4. Rank selection under serial correlation

The §3.1–§3.2 analysis takes the factor rank r as given. In practice r is selected by cross-validation, and under serial correlation the default CV selects a rank that is too large—the upstream of the residual-shrinkage mechanism analyzed in §2.1. This subsection takes up the CV question directly: how the default block design behaves under serial correlation, what recent Monte Carlo work shows, and the corresponding default change in `fect` v2.3.0.

Why block CV over-selects. The cross-validation default in `gsynth` (through v1.3.0) and `fect` (through v2.2.x) is *block CV*: each fold holds out a contiguous time block from every control unit and scores the held-out MSPE; the rank minimizing average MSPE across folds wins. The block design, together with the optional donut exclusion, was introduced precisely because held-out cells in a time-series panel are not independent of their neighbors: the contiguous block keeps each fold's training and validation regions structurally separate, and the donut absorbs short-range residual correlation across the boundary. What recent Monte Carlo work documents—and what motivates the v2.3.0 change—is that these safeguards are not enough when serial correlation is strong enough to span the donut. The residual at the held-out time t^* is then correlated with residuals at training cells outside the donut, and an over-fit factor model that absorbs the shared time-variation reduces the residuals at those training cells and, through the same correlation, reduces the predicted residual at t^* itself. Block CV then tends to choose a higher rank than the truth. The effect is symmetric across GSC and IFE-EM since it acts at the factor-fitting stage. GSC's predictive bootstrap (§3.2) absorbs the resulting in-sample residual shrinkage—Step 1 measures uncertainty against a factor basis the held-out unit did not help estimate—but the pre-v1.3.1 IFE-EM bootstrap does not (§3.1). The CV issue and the bootstrap issue are independent; the practical damage requires their combination.

Rolling-window CV as remedy. To address this, `fect` v2.3.0 adopts rolling-window CV, the standard cross-validation design for time-series problems. For each held-out unit, the procedure picks a random time point t^*, scores the model's prediction over the next short window of cells, and refits the model after dropping everything from t^* onward—the refit therefore never sees the future of its own validation window. A short buffer of g cells immediately before t^* extends the cut backward in time, absorbing any residual correlation that would otherwise carry across it. Figure 2 compares the two designs. One practical wrinkle: holding out every eligible unit at once can leave the model with too few donor cells

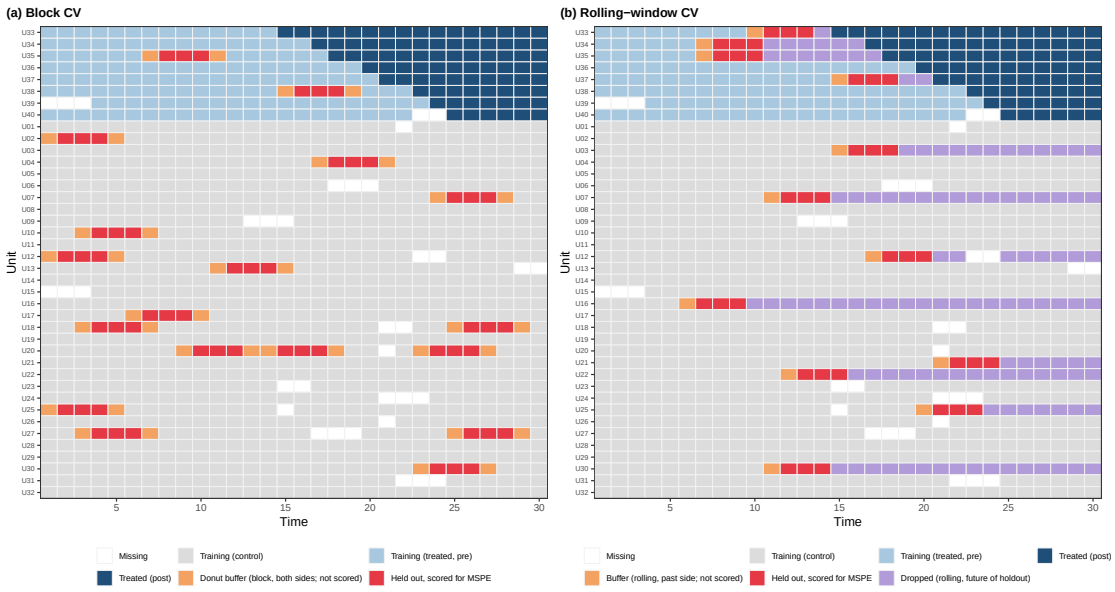


Figure 2: Block CV (the default through `fect` v2.2.x) versus rolling-window CV (the default in v2.3.0). Bright red cells are held out and scored for MSPE; sandy orange cells are buffer cells (`cv.donut` for block, `cv.buffer` for rolling); under rolling CV, light purple cells are dropped from training by the future-of-anchor cut. Block CV keeps training cells at $t^* \pm 1$ in the fit, so when residuals are serially correlated an over-fit factor model that absorbs the correlation reduces the predicted residual at the held-out cell; rolling CV separates training from validation in time and removes that mechanism.

at the held-out time points, especially on dense panels with short pre-treatment windows. `fect` therefore masks only a small fraction of eligible units per fold (the default is `cv.prop = 0.10`), with the rest contributing training data throughout. GSC and IFE-EM use the same design.

Monte Carlo evidence. Table 3 reports rank-selection results on the Xu (2017) $r_{\text{true}} = 2$ DGP at three error processes: i.i.d., AR(1) at $\rho = 0.3$, and AR(1) at $\rho = 0.8$.[3] Under i.i.d. errors both designs recover the true rank in close to all replications, with rolling CV slightly ahead (98% vs. 88%). At $\rho = 0.3$, rolling holds at 99% while block stays near 89%: the gap is small but already present once the errors carry any serial structure. At $\rho = 0.8$, the contrast is sharp: rolling recovers r_{true} in 68% of replications (mostly due to under-selection) against

[3]Shared DGP: $N_{\text{co}} = 100$, $N_{\text{tr}} = 9$, $T = 30$, $r_{\max} = 5$; 200 replications per estimator per cell (400 across estimators). Block CV runs each estimator at its pre-v2.3.0 default—`cv.method = "all_units"` for IFE-EM, `cv.method = "treated_units"` for GSC, both at `cv.nobs = 3`, `cv.donut = 0`, `cv.prop = 0.1`, `cv.rule = "1pct"`. Rolling CV runs both estimators at the v2.3.0 defaults: `cv.nobs = 3`, `cv.buffer = 1`, $k = 20$, `cv.prop = 0.10`, `cv.rule = "1se"`.

block CV's 15%; block CV pegs at $r_{\max} = 5$ in 15% of replications and selects $r \geq 3$ in 76%, while rolling never pegs and selects $r \geq 3$ in only 3%. A real-data sweep on the four Eibl and Hertog (2023) outcomes shows the same pattern: across 48 configurations spanning the buffer width and number of folds, rolling CV never pegs at the grid maximum, while block CV pegs at $r = 5$ on every cell.

	Block CV (default $\leq$ v2.2.x)			Rolling CV (default v2.3.0)		
Error process	$\bar{r}_{\text{cv}}$	% at r_{true}	% at $r_{\max}$	$\bar{r}_{\text{cv}}$	% at r_{true}	% at $r_{\max}$
i.i.d.	1.88	88	0	1.98	98	0
AR(1) $\rho = 0.3$	1.94	89	1	1.99	99	0
AR(1) $\rho = 0.8$	3.34	15	15	1.71	68	0

Table 3: Rank selection on the Xu (2017) $r_{\text{true}} = 2$ factor DGP, comparing block CV (default through v2.2.x) against rolling-window CV (default v2.3.0). $\bar{r}_{\text{cv}}$ is the mean CV-selected rank across replications; "% at r_{true}" is the share of replications selecting the true rank; "% at $r_{\max}$" is the share pegging at the grid upper bound $r_{\max} = 5$. 200 replications per estimator per cell, aggregated across GSC and IFE-EM. At AR(1) $\rho = 0.8$, block CV selects $r \geq 3$ in 76% of replications, rolling CV in 3%.

Long-range correlation is a partial exception. The G&A "Toy" DGP has Gaussian-kernel error covariance with decay scale $\approx$600, so correlation persists across the full pre-treatment window; rolling CV's anchor-and-cut design separates training from validation at the cut but cannot address correlation that spans the full panel. Widening block CV's training window similarly fails on this DGP (results not shown). When the error process has this kind of long-range structure, the placebo-based rank diagnostic of §4 remains the practical safeguard, and the first-order remedy lies at the modeling stage—a trend-cycle or harmonic pre-fit in the spirit of Shi et al. (2025) and Liu and Xu (2026)—rather than in further CV tuning.

Package response. The `gsynth/fect` codebase has been restructured around the failures analyzed in §3, with additional diagnostic and remedy work for the loading-overlap concern surfaced in the reanalysis (§4):

- *API unification* (`gsynth` v1.3.1, `fect` v2.0.0). `gsynth` v1.3.1 is now a wrapper around `fect` v2.x. The unified API parametrizes the structural choice via

`time.component.from`—`"nevertreated"` corresponds to GSC, `"notyettreated"` corresponds to IFE-EM—and the SE procedure independently via `vartype` (`"parametric"`, `"bootstrap"`, `"jackknife"`).

- *Hard gate* (`fect` v2.2.0). The bootstrap failure of §3.1 sits at one parameter intersection: `vartype="parametric"` with `time.component.from="notyettreated"`. As of v2.2.0 that combination errors on construction, with a migration message pointing the user to `vartype="bootstrap"` or `vartype="jackknife"`.
- *Default CV update* (`fect` v2.3.0). The default cross-validation is now rolling-window CV (`cv.method="rolling"`) with `cv.buffer = 1`; the previous block CV remains accessible via `cv.method="block"` for replication. This addresses the rank-inflation pathology of §3.4.
- *Overlap diagnostic and remedy* (`fect` v2.3.0). (a) The `plot()` method with `type = "loading.overlap"` on a fitted object now produces a loading-overlap diagnostic, scattering treated and control unit loadings in the first two factor dimensions with the convex hull of control loadings shaded. (b) An optional *bounded-loading* variant of GSC's loading projection restricts each treated unit's projected loading to lie within that hull.

The full refactor logic, additional diagnostic guardrails under consideration, alternatives weighed and rejected, and migration details are in Appendix A.5.

4. Reanalysis of the Empirical Applications

I replicate G&A's reanalysis of three *American Political Science Review* papers using the successor package `fect`: Gilens et al. (2021), Alsaadi (2025), and Eibl and Hertog (2023). The recommended fix—GSC with the Xu (2017) parametric bootstrap—broadly corroborates G&A's substantive conclusions, with two partial exceptions: the tort-law outcome in Gilens

et al. (2021) (§4.1) and health equity in Eibl and Hertog (2023) at the preferred rank (§4.3). The cross-paper picture, summarized at the end of this section, is that the bug's effect varies across the three datasets depending on whether the original CV-selected rank was high enough. For all three papers, we compare four bootstrap procedures:

- *GSC* (Algorithm 1): the GSC estimator with the Xu (2017) parametric bootstrap. §3.2 argues this remains appropriate.
- *Variant (i)* (Algorithm 3): the pre-v1.3.1 `gsynth` IFE-EM bootstrap (in-sample residuals, no LOO step). The implementation error G&A document.
- *Variant (ii)* (Algorithm 4): G&A's proposed fix—restoring the LOO step under IFE-EM. §3.1 argues this is necessary but not sufficient.
- *Variant (iii)* (Algorithm 5): an exploratory split-residuals IFE-EM bootstrap; included for completeness rather than as a recommended procedure.

Per-paper specifications, complete results, and diagnostic plots are in Appendix A.4. Each subsection below presents the GSC finding at the preferred rank in the main forest plot, with an appendix forest at the original pick for direct comparison with G&A's reanalysis.

Preferred-rank selection. Throughout this section I report results at what I call the *preferred rank*: the smallest r at which the event-study pre-trend gives no visible sign of time-varying confounding, with the placebo test as a supplementary check. The choice is read off the pre-treatment plot, not the post-treatment ATT—it is independent of the size or significance of the headline effect. Unlike CV, which optimizes prediction over the whole panel and ignores treatment status, this rule incorporates the treatment indicator by asking from the residual pre-treatment variation whether there are clear signs of time-varying confounding. The rule is a diagnostic rather than a formal selection procedure, and the placebo test has limited power on the panel sizes typical of these applications.

4.1. Gilens, Patterson, and Haines (2021)

The analysis in Gilens et al. (2021) of *Citizens United* examines five state-level outcomes chosen to span a gradient of corporate stake: top corporate income tax (broadest stake), tort law (moderate), eminent domain (weak), and abortion and gun control (no corporate stake). The paper's substantive claim rests on the resulting contrast—significant pro-corporate shifts on the corporate-stake outcomes (tax and tort), no detected shifts on the others—as evidence that the channel is corporate influence rather than a generic Republican-or-conservative ideological swing. G&A reanalyze all five and conclude that neither of the two originally significant findings survives a corrected inferential procedure. Rolling-window CV under `fect` v2.3.0 selects $r_{\text{cv}} \in \{0, 1, 2\}$ across the five cells; the rank-sensitivity grid (Appendix A.4.1) reports main and placebo estimates at $r \in \{0, 1, 2\}$ for each outcome, and Figure 3 shows the bootstrap-variant comparison at the preferred rank per outcome.

Top corporate income tax. At the CV-selected rank $r = 1$, $\widehat{\text{ATT}} = -2.89$ with parametric-bootstrap SE $= 2.21$ and $p = 0.19$; the placebo test at $r = 1$ also does not reject ($p = 0.19$). At $r = 0$ the ATT is significant ($p = 0.02$) but the placebo is borderline ($p = 0.06$); at $r = 2$ the bootstrap SE blows up and both the ATT and the placebo are insignificant. At every placebo-defensible specification the ATT is null. We agree with G&A: the original top-corporate-income-tax finding does not survive a valid inferential procedure.

Tort law. At the preferred $r = 2$, $\widehat{\text{ATT}} = -3.24$ ($p = 0.003$); the placebo test does not reject ($p = 0.66$), and the verdict survives across placebo-defensible ranks. Here, we register a small disagreement with G&A on this outcome. That said, there is a caveat on the statistically significant result: `civil100` is close to a step function—most states adjust their tort regime within a 2001–2005 window and remain flat thereafter—so the factor model absorbs nearly all within-unit variation into its leading factors and the residual autocorrelation collapses to near zero. The bootstrap SE is based on a residual pool maybe too small to reflect the

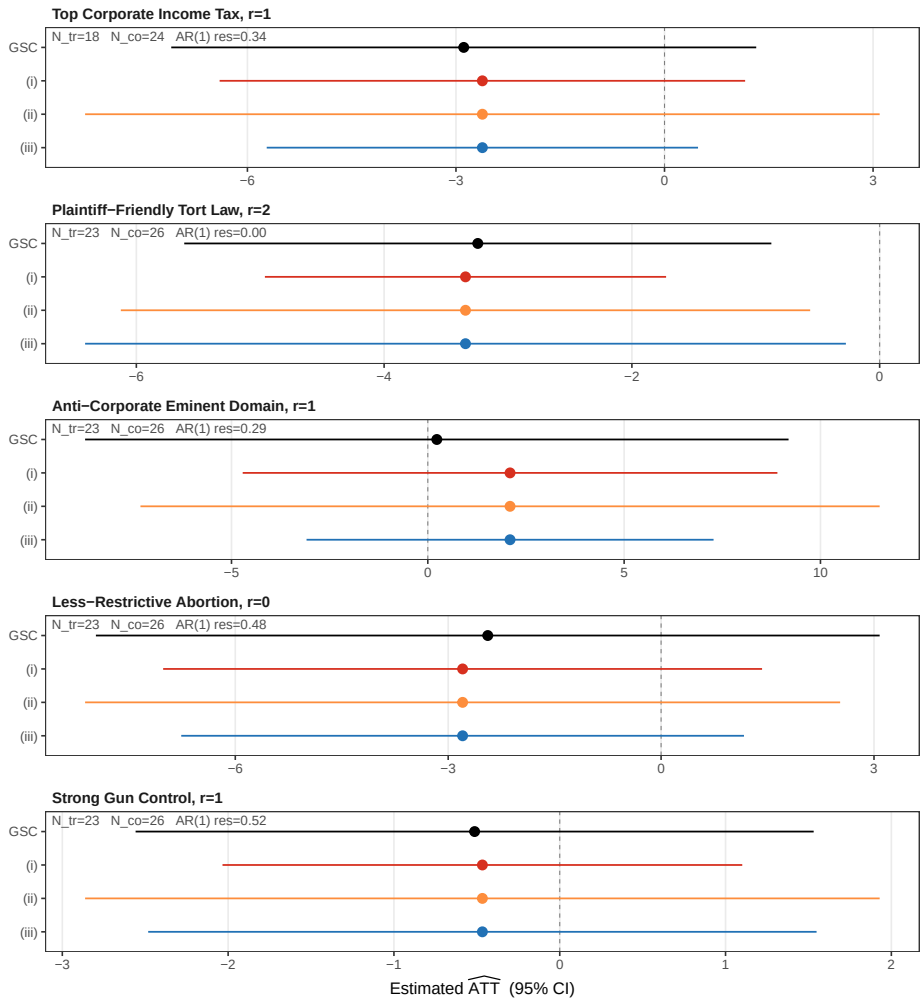
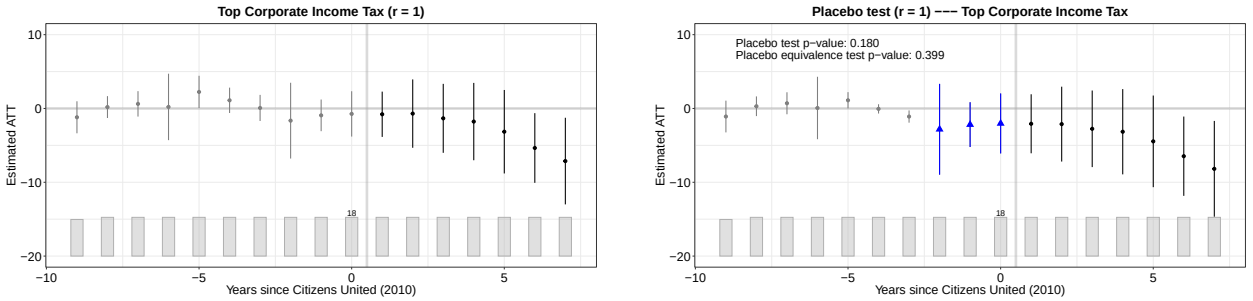


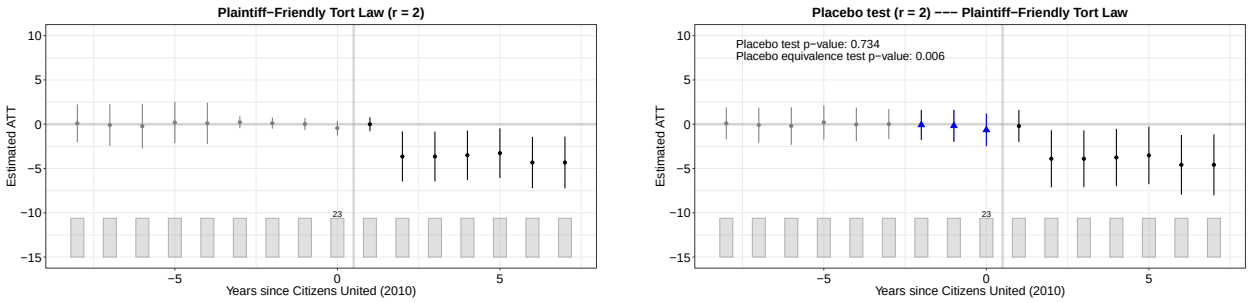
Figure 3: Point estimates and 95% confidence intervals at the preferred rank per outcome for the five outcomes in Gilens et al. (2021), under GSC with the Xu (2017) parametric bootstrap and IFE-EM Variants (i), (ii), (iii). Outcomes ordered by hypothesized corporate stake (broadest at top). Footers report N_{tr}, N_{co}, and post-fit residual autocorrelation. $B = 200$. Per-cell diagnostics are in Appendix A.4.1; Appendix A.4.1 shows the corresponding forest at the ranks the original paper picked (also what G&A reanalyzed against), with substantively the same picture.

underlying noise scale. Importantly, the parametric bootstrap relies on residuals carrying the dominant noise scale, an assumption a step-function-like outcome violates.

Abortion, gun control, and eminent domain. These three outcomes were already insignificant in the original paper, and remain insignificant under our analysis at both the CV-selected rank ($r_{\text{cv}} = 0$ for all three) and the placebo-defensible alternatives. We agree with the original authors and with G&A. Eminent domain at $r = 2$ is the one cell where



(a) Top corporate income tax at $r = 1$: gap (left) and placebo (right). $\widehat{\text{ATT}} = -2.89$ ($p = 0.19$); placebo $p = 0.19$.



(b) Tort law at $r = 2$: gap (left) and placebo (right). $\widehat{\text{ATT}} = -3.24$ ($p = 0.003$); placebo $p = 0.66$.

Figure 4: GSC gap and placebo plots at the preferred rank for the two outcomes Gilens et al. (2021) originally report as significant. Tax loses significance at any placebo-defensible rank; tort survives but with the structural caveat below. In the placebo panels, blue triangles mark the three placebo-period ATT estimates (periods -2, -1, 0 treated as post-treatment).

overfit shows up unambiguously: the placebo coefficient jumps to +65 ($p = 0.04$), a textbook overfit signature that disappears at $r \in \{0, 1\}$. Appendix A.4.1 reports the rank-sensitivity grid for each outcome.

Summary. Of the paper's two originally significant findings, tax fails under valid inference and tort survives only at the preferred rank, with the caveat above. The paper's corporate-stake contrast is partially preserved: the surviving significant finding is on a corporate-stake outcome, while the non-stake outcomes remain null.

4.2. Alsaadi (2025)

The GSC analysis in Alsaadi (2025) is one of three quantitative legs in the paper: a survival analysis on regime durability (1900–2015), OLS and logistic regressions on challenge outcomes

from the NAVCO 2.1 dataset (1945–2013), and the GSC analysis examined here. The GSC piece tests a contrast on a single continuous outcome—"mobilization for autocracy", the V-Dem index `v2caautmob_osp`—using two treatment specifications: Treatment (a) `minority` (a single majority excluded), and Treatment (b) `frac_minority` (other minorities excluded). The original paper reports a significant pro-mobilization effect under (a) and a null under (b), and reads the (a)-vs-(b) contrast as evidence that the minority-type configuration matters beyond minority status alone.[4] Under the paper's rank grid $r \in [1, 5]$, block CV (the pre-v2.3.0 default) pegs at $r = 5$ for both cells, and the residual autocorrelation after the $r = 5$ fit remains near 0.6—signals of rank over-selection. Rolling-window CV under `fect` v2.3.0 picks $r_{\text{cv}} = 3$ for (a) and $r_{\text{cv}} = 0$ for (b), no longer pegging.

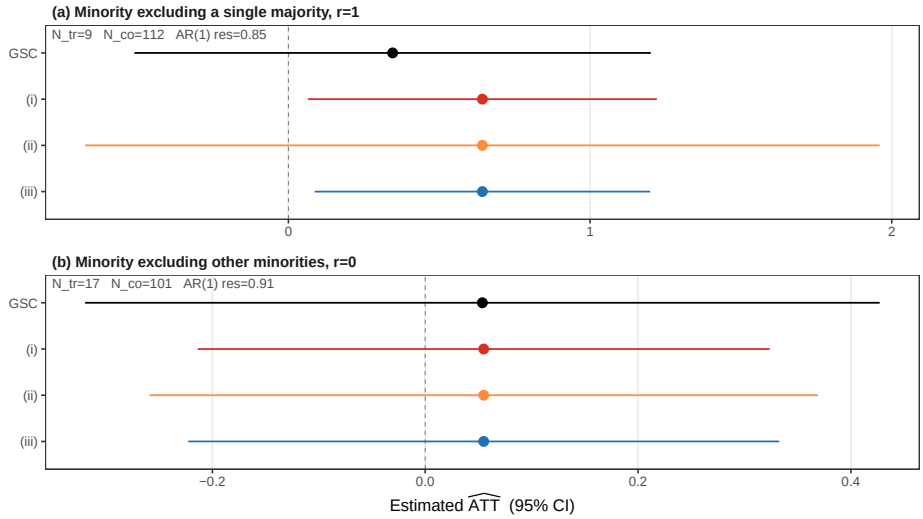
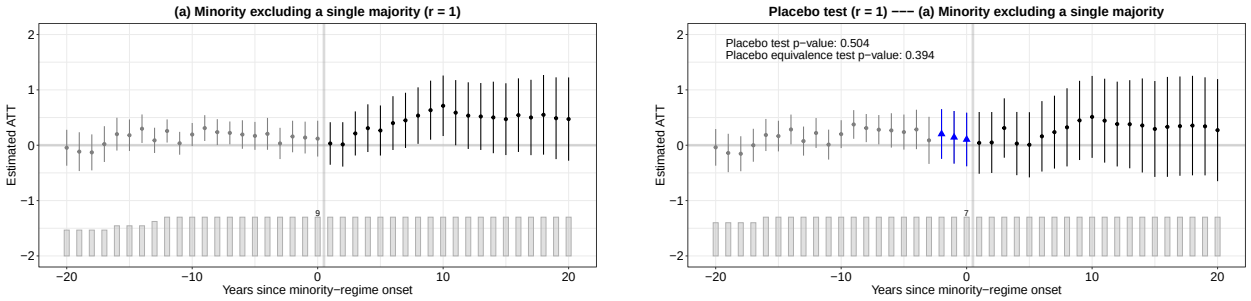


Figure 5: Point estimates and 95% confidence intervals at the preferred rank per cell ($r = 1$ for (a), $r = 0$ for (b)) for the two treatment specifications in Alsaadi (2025), under GSC with the Xu (2017) parametric bootstrap and IFE-EM Variants (i), (ii), (iii). Footers report N_{tr}, N_{co}, and post-fit residual autocorrelation. $B = 200$. Appendix A.4.2 shows the corresponding forest at the original block-CV pick ($r = 5$).

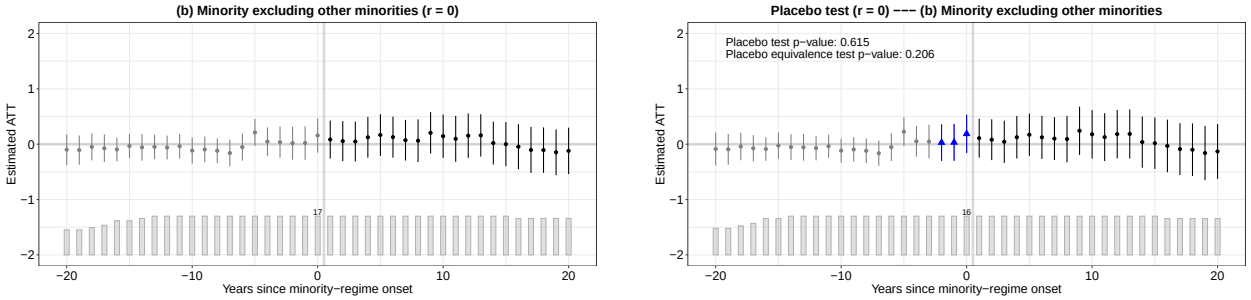
At the preferred rank for each cell, $r = 1$ for (a) and $r = 0$ for (b), GSC returns no significant effect on either cell, and the (a)-vs-(b) contrast the original paper reports no longer holds (Figure 5). Even at the original block-CV pick of $r = 5$, switching to the GSC

[4]Treatment is monotonized per country (cumulative max within country) to convert the few reversal cases into the absorbing-onset frame the GSC pipeline requires; this drops N_{tr} from the paper's 12 and 24 to 9 and 17 for (a) and (b) respectively.

estimator with parametric bootstrap is enough to render Treatment (a) not statistically significant, without invoking the preferred-rank diagnostic (Appendix A.4.2). This is distinct from G&A's proposal to add a leave-one-out step to the IFE-EM bootstrap, which §3.1 argues is necessary but not sufficient under IFE-EM. Figure 6 shows the gap and placebo plots at the preferred ranks; Appendix A.4.2 runs the GSC sensitivity grid across $r \in \{0, 1, 2, 5\}$ for each treatment.



(a) Treatment (a) `minority` at $r = 1$: gap (left) and placebo (right).



(b) Treatment (b) `frac_minority` at $r = 0$: gap (left) and placebo (right).

Figure 6: Preferred-specification gap and placebo plots for the two Alsaadi (2025) treatments. $r = 1$ is the minimum rank at which Treatment (a) clears the placebo test; $r = 0$ (imputation with unit fixed effects only, matching the original paper's specification) suffices for Treatment (b). In the placebo panels, blue triangles mark the three placebo-period ATT estimates (periods -2, -1, 0 treated as post-treatment).

4.3. Eibl and Hertog (2023)

The Eibl and Hertog (2023) analysis tests a 2×2 contrast (treatment type $\times$ country type) across four welfare outcomes: health and education equality, and primary and secondary enrollment. The original argument hinges on a single cell—oil-rich countries facing center-seeking subversion—showing systematic positive effects across all four outcomes, while the

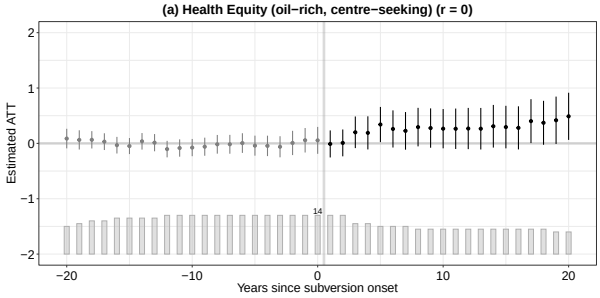
other three cells (oil-rich + separatist; oil-poor + center-seeking; oil-poor + separatist) do not. The paper supports this contrast through GSC alongside difference-in-differences as a parallel robustness leg. Both G&A and the analysis below reanalyze only the GSC piece on this headline cell; the other three cells and the DID leg are out of scope.

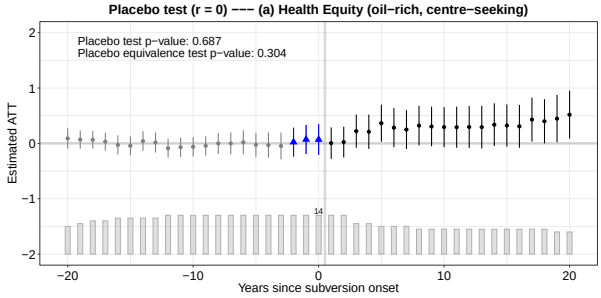
At the preferred rank for each outcome, only health equity recovers a significant positive ATT. Education equity, primary enrollment, and secondary enrollment yield point estimates in the original direction but with confidence intervals that cross zero. Of the four originally significant within-cell findings, only health equity survives at the preferred rank; the other three no longer hold. We mostly agree with G&A's reversal verdict (three of four), and register health equity as the narrow exception. The within-cell pattern that the original reports across all four welfare outcomes is materially weakened.

Block CV pegs the rank at $r = 5$ on all four outcomes, the top of the grid Eibl and Hertog use. The preferred ranks implied by the rank-sensitivity grid are much lower: $r = 0$ for health and education equity and $r = 1$ for primary and secondary enrollment. Figure 7 shows the gap (event-study) and placebo plots at the preferred rank per outcome. Appendix A.4.3 carries the full rank-sensitivity grid; at ranks above each outcome's preferred value, signs flip and placebos reject, the classic signature of overfitting.

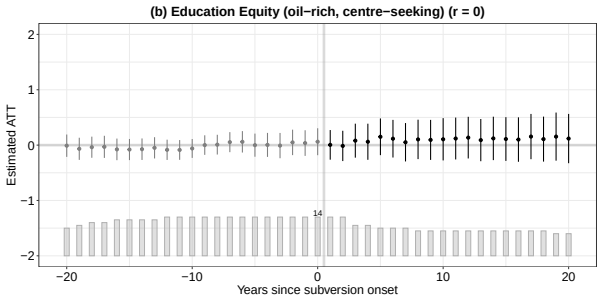
Figure 8 compares the four bootstrap procedures at each outcome's preferred rank; they broadly agree on every outcome. Here the implementation fix alone is enough: at the original $r_{\text{cv}} = 5$, switching to GSC with the parametric bootstrap renders all four within-cell findings non-significant (Appendix A.4.3). The preferred-rank diagnostic gives a softer result: health equity survives at $r = 0$. Still, both routes reach the same conclusion: the original within-cell pattern is fragile. Residual autocorrelation remains high (ranging from 0.56 to 0.88 across the four outcomes) at $r = 5$.

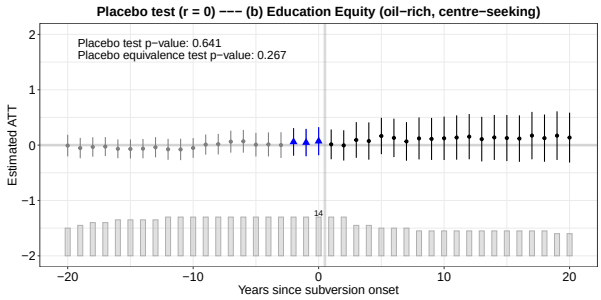
Across the three reanalyses. The effect of the bug differs across the three papers. For Gilens et al. (2021), all rank-selection rules considered choose small ranks: $r \in \{1, 2\}$ in the



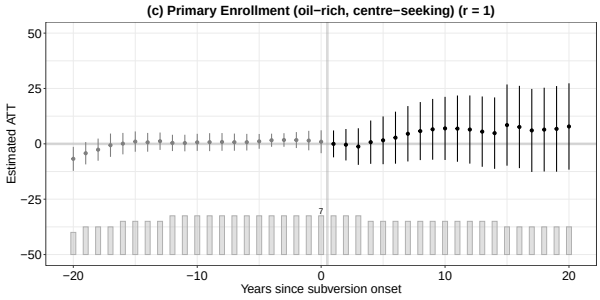


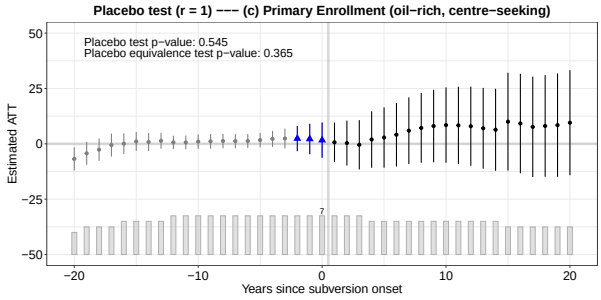
(a) Health equity at $r = 0$: gap (left) and placebo (right).



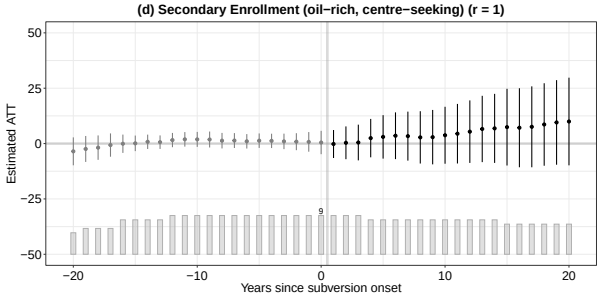


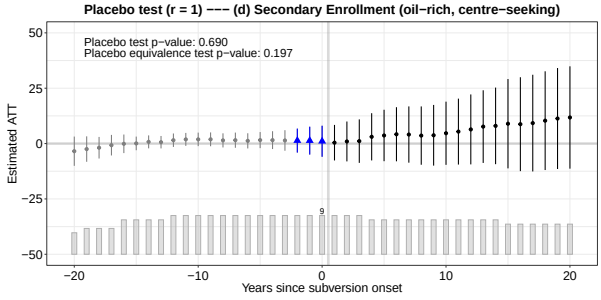
(b) Education equity at $r = 0$: gap (left) and placebo (right).





(c) Primary enrollment at $r = 1$: gap (left) and placebo (right).





(d) Secondary enrollment at $r = 1$: gap (left) and placebo (right).

Figure 7: Preferred-rank gap and placebo plots for the four Eibl and Hertog (2023) outcomes. The preferred rank is the simplest at which the placebo test does not reject, $r = 0$ for health and education equity and $r = 1$ for primary and secondary enrollment—four to five factors below the $r = 5$ CV selects for all four. At these ranks the ATT is significantly positive for every outcome. In the placebo panels, blue triangles mark the three placebo-period ATT estimates (periods -2, -1, 0 treated as post-treatment).

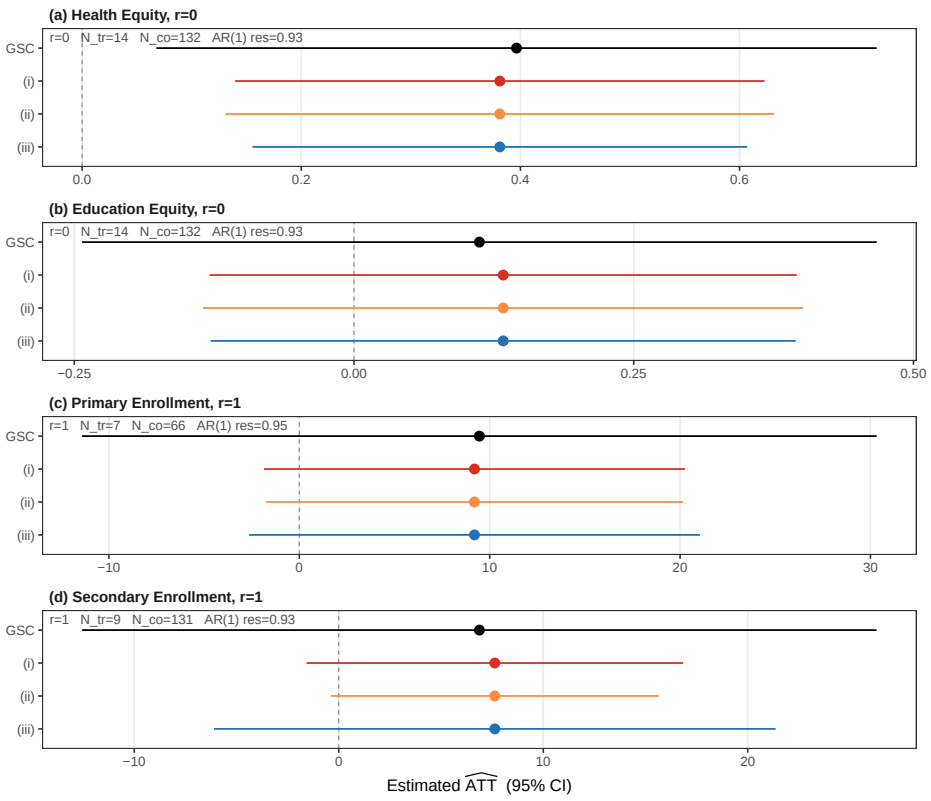


Figure 8: Inference-procedure comparison at each outcome's preferred rank for the Eibl and Hertog (2023) reanalysis: GSC (the Xu (2017) parametric bootstrap, Algorithm 1), Variant (i) (in-sample IFE-EM residuals, the pre-v1.3.1 procedure; Algorithm 3), Variant (ii) (LOO-corrected IFE-EM without sample splitting; Algorithm 4), and Variant (iii) (LOO plus control-pool split, `split_residuals=TRUE`; Algorithm 5). Points are point estimates; bars are 95% CIs. Preferred rank is the simplest at which the placebo test does not reject: $r = 0$ for health and education equity, $r = 1$ for primary and secondary enrollment. $B = 200$. Footers show the preferred rank, N_{tr}, N_{co}, and post-fit residual autocorrelation per outcome.

original paper (presumably selected by block CV) and in G&A's reanalysis, $r \in \{0, 1, 2\}$ under v2.3.0 rolling CV, and similar ranks under the pretrend-based preferred-rank rule. The data do not support high ranks in this case, so the bug has limited effect regardless of the CV rule. By contrast, for Alsaadi (2025) and Eibl and Hertog (2023), the original block-CV procedure selects $r = 5$, where the bug matters. At this rank, switching to GSC with the Xu (2017) parametric bootstrap is enough to reverse the headline finding in one of two cells for Alsaadi (2025) and in all four within-cell findings for Eibl and Hertog (2023) at $r = 5$ (three of four at the preferred rank).

G&A's leave-one-out proposal for IFE-EM reaches the same verdict via a different route

(§3.1)–Variant (ii) in this response; the two fixes converge on these applications. It is also G&A's reanalysis that surfaces the upstream rank-inflation pattern—the regime in which the bug becomes consequential in the first place—and that motivates the rolling-CV default in `fect` v2.3.0.

5. Conclusion

This note acknowledges the implementation error documented by G&A in the pre-v1.3.1 `gsynth` parametric bootstrap for IFE-EM. The note develops three points: (a) the leave-one-out correction G&A propose is necessary under GSC but not sufficient under IFE-EM; (b) the implementation passed standard i.i.d. validation through over-conservative CIs rather than correct calibration, with the failure mode appearing only under serial correlation; and (c) serially correlated data drive cross-validation to select too many factors, and this overfitting compounded with the pre-v1.3.1 inferential error leaves applied researchers overconfident about unreliable estimates.

Practical recommendations. For an applied analyst using a factor-augmented approach, especially GSC, IFE-EM, or matrix completion, for causal panel analysis, the implications are as follows.

- When N_{tr} is small but a sizable pool of never-treated units is available, the GSC estimator, in combination with the parametric bootstrap in Xu (2017), remains valid. The insulation property in §3.2 makes this bootstrap valid without additional sample splitting or distributional assumptions on the idiosyncratic error.
- When N_{tr} is large and the IFE-EM or matrix completion estimators are appealing, use a nonparametric cluster-bootstrap or jackknife instead of the parametric bootstrap. The parametric option for IFE-EM is no longer available in the current `fect` release.
- Inspect the event-study plot and conduct placebo tests. The purpose of the factor-

augmented approach is to adjust for time-varying confounding that can be represented by a low-rank structure. For this reason, prefer a parsimonious model with a smaller rank to avoid overfitting. The rank should not be chosen to obtain more significant results or smaller p-values. Be wary of rank selection that may lead to overfitting when outcomes are highly serially correlated.

- Before committing to a factor-model approach, inspect the `"loading.overlap"` diagnostic now available from `fect`'s `plot()` method. If treated loadings sit outside the convex hull of control loadings, the counterfactual is an extrapolation; `fect` also provides a bounded-loading variant of GSC's loading projection designed for this case.

A broader view. Stepping back, the failure has two sources working together. Cross-validation under serial correlation tends to select too many factors, and the pre-v1.3.1 parametric bootstrap for IFE-EM did not correct for the resulting misspecification. Together, these problems produce overfit estimates with overconfident intervals. The cases in which cross-validation over-selects factors are also the cases in which the bootstrap error matters most. Four themes emerge from the exchange that sit beyond this note's immediate scope.

- *Low-rank approximation remains useful.* GSC, IFE-EM, synthetic difference-in-differences (SDID) (Arkhangelsky et al., 2021), and matrix-completion estimators (Athey et al., 2021) all rest on the assumption that the control panel admits a small-rank representation of the outcome residual matrix. The assumption is not innocuous, but it appears to have held up well in practice: many panel outcomes *are* well-approximated by a few latent time-varying drivers, which is why these methods coexist productively in applied work. Recent estimators continue to extend this framework—for example, Athey et al. (2026) combine a low-rank outcome model with the unit and time weights of SDID, suggesting that an explicit factor structure carries information beyond what unit/time weighting alone supplies. The implementation error corrected here does not touch the low-rank assumption itself; it concerns how uncertainty is quantified on top of it.

- *Overlap and the simplex constraint.* When treated-unit loadings sit outside the convex hull of control loadings, making the estimated counterfactual purely an extrapolation based on the factor model. The classical synthetic control method (Abadie et al., 2010) imposes a simplex constraint on weights that mechanically rules out this extrapolation; SDID retains the same simplex constraint on control-unit weights and add an L_2 penalty plus a free intercept term to relax the pre-trend matching requirement. GSC's loading-projection step has historically been unconstrained OLS, which permits negative or out-of-hull weights. The latest version of `fect` (v2.3.0) adds a *bounded-loading* algorithm that restricts the projected loading for each treated unit to the convex hull of control loadings, supplying a built-in overlap safeguard in the spirit of classical synthetic control while preserving the factor-model structure.

- *Rank selection under serial correlation.* When residuals are strongly serially correlated, block cross-validation tends to select too many factors. Adjacent training and held-out cells share residual structure across the cut, so a model that overfits training-cell variation can also fit the held-out cell well (§3.4). The `fect` v2.3.0 default therefore switches to rolling-window CV. This design cuts the training panel forward in time at a random anchor and drops a buffer of cells immediately before the anchor from training. It is a standard time-series cross-validation design that reduces cross-fold leakage and substantially limits rank inflation under AR(1) errors. The remaining open case is long-range correlation, where the buffer that handles AR(1) errors cannot be widened without cost. The modeling-stage response below is the appropriate way to address that case.

- *Remedies for non-stationarity.* Classical synthetic control and factor-model panel methods both face a problem under outright non-stationarity (and non-cointegration) that CV tuning cannot solve. Shi et al. (2025) document the "spurious synthetic control problem" in non-stationary macroeconomic data: units may share trends rather than cyclical structure, leading to misleading counterfactuals. They propose separating trend and cycle before

fitting the SC. The same modeling-stage remedy applies in the factor-model setting: use a trend–cycle or harmonic pre-fit, in the spirit of Shi et al. (2025) and Liu and Xu (2026), rather than further CV tuning.

- *Predictive error and conformal inference.* The leave-one-out predictive errors driving Step 1 of Xu (2017)'s Algorithm 2 are a special case of a non-conformity score in the sense of conformal prediction. Split-conformal procedures (Lei and Candès, 2021; Angelopoulos and Bates, 2023) produce finite-sample distribution-free prediction intervals from such scores under exchangeability alone. The closest existing work in this direction is Cattaneo et al. (2021) and Cattaneo et al. (2025), which develop non-asymptotic prediction intervals for the classical synthetic control estimator under time-side regularity (i.i.d., β-mixing, or cointegrated). Ongoing work with Liu (Liu and Xu, 2026) develops a conformal recasting of Algorithm 2 for the GSC factor-model setting, replacing the residual-bootstrap justification with a finite-sample guarantee under unit-level exchangeability.

- *Sample splitting beyond the parametric case.* The argument in §3.2 that sample splitting is not needed under GSC rests on the factor model being parametric at fixed rank, with the fast convergence rates of Bai (2009). Once one moves to matrix completion with data-driven rank or to nonlinear or ML-based imputers, that parametric structure is lost and DML-style cross-fitting (Chernozhukov et al., 2018) becomes the natural framework; Abadie et al. (2024) develop a doubly robust inference procedure for causal latent factor models in this spirit. Sample splitting on the control pool—GSC+split explored in the Monte Carlo—is the lightest version of this idea, and would be the appropriate next step if the imputer is allowed to be more flexible.

These directions are incremental. None abandons the low-rank framework, and together they sharpen its inferential foundation in the cases where the current parametric machinery is most strained, such as when N_{tr} is small.

Finally, I thank Beniamino Green and P. M. Aronow for identifying the implementation

error, for documenting it carefully, and for conducting the reanalyses that show its practical stakes. I regret the error and the years it persisted. I also recognize the cost this places on the authors of the affected applications, whose published findings may now need to be revisited. The implementation error was mine. The broader lesson is that placebo and rank-sensitivity diagnostics should be routine in factor-model panel work, and that serial correlation in panel settings remains understudied and underemphasized. This lesson extends beyond the three papers reanalyzed here. If this exchange has any lasting value beyond the specific fix, I hope it sharpens the community's understanding of credible inferential methods and the risks of overfitting in applied social science research.

References

Abadie, A., Agarwal, A., Dwivedi, R., and Shah, A. (2024). Doubly robust inference in causal latent factor models. arXiv:2402.11652.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Alsaadi, S. (2025). Unconditional loyalty: The survival of minority autocracies. *American Political Science Review*, 120(1):189–206.

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix

completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730.

Athey, S., Imbens, G., Qu, Z., and Viviano, D. (2026). Triply robust panel estimators. arXiv:2508.21536.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Cattaneo, M. D., Feng, Y., Palomba, F., and Titiunik, R. (2025). Uncertainty quantification in synthetic controls with staggered treatment adoption. *Review of Economics and Statistics*. forthcoming.

Cattaneo, M. D., Feng, Y., and Titiunik, R. (2021). Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68.

Eibl, F. and Hertog, S. (2023). From rents to welfare: Why are some oil-rich states generous to their people? *American Political Science Review*, 118(3):1324–1343.

Gilens, M., Patterson, S., and Haines, P. (2021). Campaign finance regulations and public policy. *American Political Science Review*, 115(3):1074–1081.

Gobillon, L. and Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, 98(3):535–551.

Gonçalves, S. and Perron, B. (2014). Bootstrapping factor-augmented regression models. *Journal of Econometrics*, 182(1):156–173.

Green, B. and Aronow, P. M. (2026). Undocumented behavior in the gsynth r package and its consequences for three published studies. Working paper, April 13, 2026.

Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B*, 83(5):911–938.

Liu, L., Wang, Y., and Xu, Y. (2024). A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 68(1):160–176.

Liu, Z. and Xu, Y. (2026). The harmonic synthetic control method. Mimeo, UC Berkeley and Stanford.

Shi, Z., Xi, J., and Xie, H. (2025). A synthetic business cycle approach to counterfactual analysis with nonstationary macroeconomic data. *arXiv preprint arXiv:2505.22388*.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

A. Appendix

Table of Contents

A.1. The five bootstrap procedures

For self-containment, this appendix records the five bootstrap procedures evaluated in Appendix A.3 as explicit pseudocode. All five share common inputs, two factor-model subroutines, and a common output structure:

- **Inputs.** The observed panel $\{Y_{it}, X_{it}, D_{it}\}$; control set $\mathcal{C}$, treated set $\mathcal{T}$, pre-period length T_0; rank r (from cross-validation); bootstrap counts (B_1, B_2) following Xu (2017).
- **FitGSC**(S) with $S \subseteq \mathcal{C}$: returns $(\hat{F}, \hat{\Lambda}, \hat{\beta})$ from a rank-r factor-model fit on S alone; for units $i \notin S$, projects $\hat{\lambda}_i$ from the pre-T_0 data of i; returns the counterfactual $\hat{Y}_{it}(0) = X_{it}\hat{\beta} + \hat{\lambda}_i'\hat{f}_t$.
- **FitIFE_EM**(S) with $S \subseteq \mathcal{C}$: solves the joint objective over $S \cup \mathcal{T}_{\text{pre}}$ with EM imputation of $\mathcal{T}_{\text{post}}$, returning the analogous quantities.
- **Residuals (unit vectors).** For each control unit i, the in-sample residual vector $\hat{\boldsymbol{\varepsilon}}_i = (\hat{\varepsilon}_{i,1}, \ldots, \hat{\varepsilon}_{i,T})$ with $\hat{\varepsilon}_{it} = Y_{it} - \hat{Y}_{it}$. For each leave-one-out draw m, the prediction-error vector $\hat{\boldsymbol{\varepsilon}}^p_{(m)} = (\hat{\varepsilon}^p_{(m),T_0+1}, \ldots, \hat{\varepsilon}^p_{(m),T})$ over the post-period of the held-out unit. Resampling in every algorithm below is by unit (vector), not by cell, preserving within-unit serial structure.
- **Target.** ATT estimator $\widehat{\text{ATT}}_t = N_{\text{tr}}^{-1} \sum_{i \in \mathcal{T}} (Y_{it} - \hat{Y}_{it}(0))$ for $t > T_0$.

G&A reproduce Xu (2017)'s Algorithm 2 as their Algorithm 1 (p. 5) and the pre-v1.3.1 IFE-EM bootstrap as their Algorithm 2 (p. 7). Algorithm 3 below matches G&A's Algorithm 2 line-for-line. Algorithm 1 departs from Xu (2017)'s Algorithm 2 in one respect, flagged in a footnote below, which Xu (2017)'s own prose (p. 24) describes in the simpler form adopted here and which the `gsynth` code base has implemented since its first CRAN release.

Algorithm 1 is the benchmark procedure of §3.2: factors are estimated from controls only, and an LOO loop over $\mathcal{C}$ supplies honest out-of-sample prediction errors for the treated-post bootstrap perturbation.

Algorithm 1 GSC with the Xu (2017) parametric bootstrap.

Inputs. Panel $\{Y_{it}, X_{it}, D_{it}\}$; $\mathcal{C}$, $\mathcal{T}$, T_0; rank r; counts (B_1, B_2).

Step 0 (main fit). Run FitGSC($\mathcal{C}$) to obtain $(\hat{F}, \hat{\Lambda}, \hat{\beta})$, the counterfactual $\hat{Y}_{it}(0)$, and $\widehat{\text{ATT}}_t$ for $t > T_0$. Record the in-sample residual pool $\hat{\mathcal{E}} = \{\hat{\boldsymbol{\varepsilon}}_i : i \in \mathcal{C}\}$, a set of $|\mathcal{C}|$ unit residual vectors of length T.

Step 1 (out-of-sample prediction errors, LOO on $\mathcal{C}$).[A1]

(i) For $m = 1, \ldots, B_1$: draw $i^* \in \mathcal{C}$ uniformly at random; run FitGSC($\mathcal{C} \setminus \{i^*\}$); project $\hat{\lambda}_{i^*}^{(-i^*)}$ from i^*'s pre-T_0 data; record $\hat{\varepsilon}^p_{(m),t} = Y_{i^*,t} - X_{i^*,t}\hat{\beta}^{(-i^*)} - \hat{\lambda}_{i^*}^{(-i^*)\prime}\hat{f}_t^{(-i^*)}$ for $t > T_0$.

(ii) Collect the prediction-error pool $\hat{\mathcal{E}}^p = \{\hat{\boldsymbol{\varepsilon}}^p_{(m)} : m = 1, \ldots, B_1\}$, a set of B_1 vectors of length $T - T_0$.

Step 2 (bootstrap replications).

(i) For $k = 1, \ldots, B_2$:

- For each control unit $i \in \mathcal{C}$: draw a unit index $j \sim \text{Unif}(\mathcal{C})$ and set $Y^{(k)}_{i,\cdot} = \hat{Y}_{i,\cdot} + \hat{\boldsymbol{\varepsilon}}_j$ (the full T-vector of unit j's in-sample residuals).
- For each treated unit $i \in \mathcal{T}$: draw $m \sim \text{Unif}\{1, \ldots, B_1\}$ and set $Y^{(k)}_{it} = \hat{Y}_{it}(0) + \hat{\varepsilon}^p_{(m),t}$ for $t > T_0$; $Y^{(k)}_{it} = Y_{it}$ for $t \le T_0$.

(ii) Run FitGSC($\mathcal{C}$) on $\{Y^{(k)}_{it}\}$; record $\widehat{\text{ATT}}^{(k)}_t$.

Return. $\widehat{\text{ATT}}_t$, $\widehat{\text{SE}}_t = \sqrt{\text{Var}_k\{\widehat{\text{ATT}}^{(k)}_t\}}$, and the 95% confidence interval by the normal wrapper $\widehat{\text{ATT}}_t \pm 1.96 \cdot \widehat{\text{SE}}_t$ (Xu (2017) prescribes the percentile method in Step 4 of Algorithm 2; `gsynth` switched to the normal wrapper in v1.1.7, discussed in §3.2).

Algorithm 2 replaces the LOO loop with a one-shot split of $\mathcal{C}$; it is included as a diagnostic contrast, not a recommendation, since Table A1 shows splitting adds no coverage benefit under GSC's insulation property.

Algorithm 2 GSC with control-pool sample splitting.

Inputs. As in Algorithm 1. Partition $\mathcal{C} = \mathcal{C}_A \sqcup \mathcal{C}_B$ uniformly at random once.

Step 0 (main fit). Run FitGSC($\mathcal{C}_A$) to obtain $(\hat{F}^{(A)}, \hat{\Lambda}^{(A)}, \hat{\beta}^{(A)})$, the counterfactual $\hat{Y}_{it}(0)$, and $\widehat{\text{ATT}}_t$.

Step 1 (out-of-sample prediction errors, via the split). $\mathcal{C}_B$ is out-of-sample with respect to Step 0, so no LOO inner loop is needed. For each unit $i \in \mathcal{C}_B$, define the prediction-error vector $\hat{\boldsymbol{\varepsilon}}_i^p = (\hat{\varepsilon}_{i,1}^p, \ldots, \hat{\varepsilon}_{i,T}^p)$ with $\hat{\varepsilon}_{it}^p = Y_{it} - \hat{Y}_{it}(0)$, and collect the pool $\hat{\mathcal{E}}^p = \{\hat{\boldsymbol{\varepsilon}}_i^p : i \in \mathcal{C}_B\}$.

Step 2 (bootstrap replications).

(i) For $k = 1, \ldots, B_2$: for each unit $i \in \mathcal{C} \cup \mathcal{T}$, draw a vector $\hat{\boldsymbol{\varepsilon}}_i^* \sim \text{Unif}(\hat{\mathcal{E}}^p)$ and set $Y_{i,\cdot}^{(k)} = \hat{Y}_{i,\cdot}(0) + \hat{\boldsymbol{\varepsilon}}_i^*$ on cells indexed by $\mathcal{C}$ and on the post-period of $\mathcal{T}$; for $i \in \mathcal{T}$, $t \le T_0$ set $Y_{it}^{(k)} = Y_{it}$.

(ii) Run FitGSC($\mathcal{C}_A$) on $\{Y_{it}^{(k)}\}$ (same partition); record $\widehat{\text{ATT}}_t^{(k)}$.

Return. As in Algorithm 1.

Algorithm 3 is the pre-v1.3.1 `gsynth` procedure for IFE-EM: Step 1 is absent entirely, and both control-cell and treated-post bootstrap perturbations are drawn from the in-sample residual pool—the error G&A document.

Algorithm 3 IFE-EM, Variant (i) (pre-v1.3.1 `gsynth`).

Inputs. As in Algorithm 1.

Step 0 (main fit). Run FitIFE_EM($\mathcal{C}$): jointly estimate $(\hat{F}, \hat{\Lambda}, \hat{\beta})$ on $\mathcal{C} \cup \mathcal{T}_{\text{pre}}$ with EM imputation of $\mathcal{T}_{\text{post}}$; obtain $\hat{Y}_{it}(0)$ and $\widehat{\text{ATT}}_t$. Record the in-sample residual pool $\hat{\mathcal{E}} = \{\hat{\boldsymbol{\varepsilon}}_i : i \in \mathcal{C}\}$ of unit residual vectors.

Step 1. *Omitted.* No out-of-sample prediction-error pool is constructed.

Step 2 (bootstrap replications).

(i) For $k = 1, \ldots, B_2$: for each unit $i \in \mathcal{C}$, draw $j \sim \text{Unif}(\mathcal{C})$ and set $Y_{i,\cdot}^{(k)} = \hat{Y}_{i,\cdot} + \hat{\boldsymbol{\varepsilon}}_j$. For each treated unit $i \in \mathcal{T}$, draw $j \sim \text{Unif}(\mathcal{C})$ and set $Y_{it}^{(k)} = \hat{Y}_{it}(0) + \hat{\varepsilon}_{j,t}$ for $t > T_0$; $Y_{it}^{(k)} = Y_{it}$ for $t \le T_0$.

(ii) Run FitIFE_EM($\mathcal{C}$) on $\{Y_{it}^{(k)}\}$; record $\widehat{\text{ATT}}_t^{(k)}$.

Return. As in Algorithm 1.

Algorithm 4 is G&A's proposed fix: Step 1 of Algorithm 1 is restored with IFE-EM as the fitter, which §3.1 argues is necessary but insufficient, since the two contamination channels in

the IFE-EM re-fit remain.

Algorithm 4 IFE-EM, Variant (ii) (G&A's LOO fix).

Inputs. As in Algorithm 1.

Step 0 (main fit). As in Algorithm 3.

Step 1 (out-of-sample prediction errors, LOO on $\mathcal{C}$).

(i) For $m = 1, \ldots, B_1$: draw $i^* \in \mathcal{C}$ uniformly at random; run FitIFE_EM($\mathcal{C} \setminus \{i^*\}$) (the objective still includes $\mathcal{T}_{\text{pre}}$ and still imputes $\mathcal{T}_{\text{post}}$); project $\hat{\lambda}_{i^*}^{(-i^*)}$ from i^*'s pre-T_0 data; record $\hat{\varepsilon}^p_{(m),t}$ for $t > T_0$.

(ii) Collect $\hat{\boldsymbol{\varepsilon}}^p$ as in Algorithm 1.

Step 2 (bootstrap replications). As in Algorithm 1, with every call to FitGSC($\mathcal{C}$) replaced by FitIFE_EM($\mathcal{C}$).

Return. As in Algorithm 1.

Algorithm 5 grafts the split of Algorithm 2 onto IFE-EM; Appendix A.2.3 shows this closes the control-side contamination channel but leaves the treated-pre channel open, explaining the partial coverage recovery in Table A1.

Algorithm 5 IFE-EM, Variant (iii) (LOO + control-pool split).

Inputs. As in Algorithm 1. Partition $\mathcal{C} = \mathcal{C}_A \sqcup \mathcal{C}_B$ uniformly at random once.

Step 0 (main fit). Run FitIFE_EM($\mathcal{C}_A$): objective over $\mathcal{C}_A \cup \mathcal{T}_{\text{pre}}$ with EM imputation of $\mathcal{T}_{\text{post}}$. Obtain $\hat{Y}_{it}(0)$ and $\widehat{\text{ATT}}_t$.

Step 1 (out-of-sample prediction errors, via the split). $\mathcal{C}_B$ is out-of-sample with respect to Step 0, so no LOO inner loop is needed. Collect $\hat{\boldsymbol{\varepsilon}}^p$ as in Algorithm 2.

Step 2 (bootstrap replications). As in Algorithm 2, with every call to FitGSC($\mathcal{C}_A$) replaced by FitIFE_EM($\mathcal{C}_A$).

Return. As in Algorithm 1.

A.2. Extended analysis of the IFE-EM leave-one-out bootstrap

This section extends the two-issue account of §3.1. The first two subsubsections work through the arguments behind the prose sketches there: the covariance structure of the held-out residuals under the contaminated LOO refit, and the rank-overfit shrinkage mechanism together with the predictive-based reason Algorithm 1 escapes it while IFE-EM does not. The remaining three subsubsections address partial fixes—sample splitting on the control pool (§A.2.3)—alternatives considered and rejected (§A.2.4), and the large-N_{tr} regime. The treatment throughout is informal: decompositions and leading-order arguments meant to make the prose claims of §3.1 concrete enough to check, not formal proofs. The cumulative conclusion is that no modification of the parametric bootstrap rescues IFE-EM without being dominated by the nonparametric bootstrap or jackknife already available in `fect`.

A.2.1. Treated-pre contamination

Equation (7) in the main text writes out the leave-one-out re-fit's objective under IFE-EM. This subsection works out the covariance structure underlying that objective. Consider two control units $i, j \in \mathcal{C} \setminus \{i^*\}$. The held-out residuals from the IFE-EM leave-one-out have the covariance decomposition

$$\operatorname{Cov}\left(\hat{\varepsilon}_{it}^{p,\text{IFE}}, \hat{\varepsilon}_{js}^{p,\text{IFE}}\right) \;=\; \mathbf{1}\{i=j\}\cdot\sigma_{ts}^2 \;+\; \overline{\lambda}^{\top}\operatorname{Cov}\left(\hat{f}_t^{(-i^*)}, \hat{f}_s^{(-i^*)}\right)\overline{\lambda} \;+\; O(N^{-1}), \tag{A1}$$

where $\sigma_{ts}^2 \equiv \operatorname{Cov}(\varepsilon_{it}, \varepsilon_{is})$ is the noise auto-covariance, and the second term captures joint uncertainty in the factor estimate. Under GSC, the factor estimate is a function of $\mathcal{C} \setminus \{i^*\}$ alone, so the second term shrinks at $O(N_{\text{co}}^{-1})$ and is recovered asymptotically by a bootstrap that resamples controls. Under IFE-EM, the factor estimate is a function of $\mathcal{C} \setminus \{i^*\}$ *and* $\mathcal{T}_{\text{pre}}$, so the second term includes variation contributed by $\mathcal{T}_{\text{pre}}$—and the bootstrap Loop 2, which re-uses the same $\mathcal{T}_{\text{pre}}$ cells across replications, does not integrate over this contribution. The held-out residual fails the out-of-sample property not as a matter of bias but as a matter of dependence structure.

A.2.2. Residual shrinkage under CV-selected rank

Factor models with a CV-selected rank overfit serially correlated data: when r is large, the estimated factors absorb the systematic time-variation of the errors and the in-sample residuals shrink toward zero. The standard least-squares identity $\frac{1}{NT}\sum_{i,t}\mathbb{E}[\hat{\varepsilon}_{it}^2] = \sigma^2(1 - d_{\text{eff}}/NT)$ captures this in cell-averaged form, with d_{eff} the effective parameter count scaling in r. The mechanism is not specific to IFE-EM; GSC's in-sample control residuals exhibit the same

shrinkage. Xu (2017)'s Algorithm 2 is robust to this form of misspecification because it is *predictive-based*: Step 1 resamples leave-one-out out-of-sample prediction errors rather than in-sample residuals, and under GSC's insulation property (§3.2) those prediction errors measure honest out-of-sample uncertainty regardless of in-sample overfit. Under IFE-EM the LOO refit is not honestly out-of-sample—the preceding argument shows that $\mathcal{T}_{\text{pre}}$ continues to shape $\hat{F}^{(-i^*)}$—so the prediction errors inherit the in-sample shrinkage rather than escaping it. The EM step additionally inflates d_{eff} beyond the nominal $(N+T)r$ by the imputed-block dimension, heuristically of order $N_{\text{tr}} \cdot (T - T_0) \cdot r / T_0$ under the factor model (a careful derivation under EM imputation is left to future work). The leave-one-out step under IFE-EM resamples from this doubly shrunken pool.

A.2.3. Sample splitting on the control pool

A natural further modification is to split the control pool $\mathcal{C}$ into two disjoint halves $\mathcal{C}_A, \mathcal{C}_B$; estimate the factor space from $\mathcal{C}_A$ alone; collect residuals from $\mathcal{C}_B$; and resample bootstrap controls from $\mathcal{C}_B$. Under IFE-EM with the split, the factor estimator solves

$$\hat{F}^{(A)} \;=\; \arg\min_{F} \min_{\{\lambda_i\},\beta} \sum_{i \in \mathcal{C}_A} \sum_{t=1}^{T} (\,\cdot\,)^2 \;+\; \sum_{i \in \mathcal{T}} \sum_{t \le T_0} (\,\cdot\,)^2, \tag{A2}$$

and the second sum—the treated-pre block—is retained regardless of how the control pool is partitioned. Splitting alters the first sum but not the second. A held-out residual on $i \in \mathcal{C}_B$ decomposes as

$$\hat{\varepsilon}_{it}^{p,\text{split}} \;=\; \varepsilon_{it} + (\lambda_i - \hat{\lambda}_i^{(A)})' \hat{f}_t^{(A)} + \overline{\lambda}' (f_t - \hat{f}_t^{(A)}) + x_{it}^{\top} (\beta - \hat{\beta}^{(A)}), \tag{A3}$$

where $\hat{f}_t^{(A)}$ depends on $\mathcal{C}_A \cup \mathcal{T}_{\text{pre}}$. The unit i is out-of-sample with respect to $\mathcal{C}_A$, but the factor space against which its residual is measured is shaped by $\mathcal{T}_{\text{pre}}$—which also shapes the bootstrap Loop 2 refits, because those refits retain the original treated-pre cells as inputs. The split closes the contamination channel on the control side and leaves it open on the treated-pre side. A coverage simulation on the G&A "Toy" DGP (Appendix A.3, Table A1) shows Variant (iii) coverage of approximately 0.89 at nominal 0.95, materially better than Variants (i) and (ii) but still materially below nominal.

A.2.4. Further corrections considered and rejected

Three additional corrections are worth naming. First, one could also split the treated sample $\mathcal{T}$ into halves $\mathcal{T}_A, \mathcal{T}_B$, using $\mathcal{T}_A$ for factor estimation and $\mathcal{T}_B$ for loading projection. This severs

the treated-pre contamination. The obstacle is sample size: in the regime that motivates the parametric bootstrap (N_{tr} small), halving $\mathcal{T}$ is not a feasible operation. Second, one could estimate the residual-shrinkage factor from the model and inflate $\hat{\varepsilon}$ before feeding it into Step 1. The obstacle is that the variance estimate used to compute the shrinkage factor is itself contaminated by the EM procedure it attempts to correct. Third, one could resample the imputed treated cells from an asymptotic normal distribution derived from large-N, large-T approximations (e.g., Bai, 2003). The obstacle is that these approximations require both $N_{\text{tr}} \to \infty$ and $T \to \infty$ at specific rates, and the small-N_{tr} regime fails the premise. Each of these three corrections deploys a tool whose own quality depends on the asymptotic regime that is failing; the corrections degrade with the severity of the problem.

A.2.5. The large-N_{tr} regime

One might ask whether, at large N_{tr}, the corrections in Appendix A.2.4 eventually succeed. They may, but the question is not material: at large N_{tr} the nonparametric bootstrap and jackknife are asymptotically valid under standard conditions (see, e.g., Liu et al., 2024, for bootstrap consistency in panel factor-model settings), are computationally cheaper than the parametric alternative, and are already implemented in `fect` via `vartype="bootstrap"` and `vartype="jackknife"`. Even if the parametric bootstrap were eventually valid for IFE-EM at large N_{tr} after one of the corrections in Appendix A.2.4, it would remain dominated by the existing alternatives. There is no regime in which the combination of IFE-EM with the parametric bootstrap is the preferred inference procedure.

A.3. Full Monte Carlo study

The headline Monte Carlo in the main text (Table 1) draws from a larger study spanning seven data-generating processes. This appendix reports the full results.

A.3.1. DGPs

The study covers seven DGPs in two groups, distinguished by the structure of the idiosyncratic errors. The *structured-errors* group has non-identity covariance in ε_{it}:

- *G&A Toy* (long-range correlation, no factors), $N_{\text{co}} = 50$. Error covariance is $K_{tt'} = 10.2 \cdot \mathbf{1}\{t = t'\} + 10 \exp\{-(t - t')^2/600\} \cdot \mathbf{1}\{t \neq t'\}$, which yields persistent low-frequency structure across the full pre-treatment window.
- *G&A Toy*, $N_{\text{co}} = 100$. Same covariance, larger control pool.
- *Xu* $r = 2$ *with* AR(1) $\rho = 0.8$*, rank correctly specified* ($r_{\text{fit}} = 2$). The Xu (2017) two-factor structure, errors $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + \eta_{it}$ with $\eta_{it} \sim \mathcal{N}(0, 1)$.
- *Xu* $r = 2$ *with* AR(1) $\rho = 0.8$*, rank over-specified* ($r_{\text{fit}} = 4$). Same DGP, fit at twice the true rank.

The *i.i.d.-errors* group uses the Xu (2017) $r = 2$ factor structure with independent $\mathcal{N}(0, 1)$ errors:

- *Xu* $r = 2$ *i.i.d., rank correctly specified* ($r_{\text{fit}} = 2$).
- *Xu* $r = 2$ *i.i.d., mildly over-specified* ($r_{\text{fit}} = 4$).
- *Xu* $r = 2$ *i.i.d., severely over-specified* ($r_{\text{fit}} = 6$).

Shared parameters across all DGPs: $N_{\text{tr}} = 5$, $T = 30$, $T_0 = 20$ (ten post-treatment periods); $N_{\text{co}} = 50$ except where noted; 200 Monte Carlo replications per cell; $B = 100$ bootstrap replicates per replication. Under the Xu (2017) DGPs the nuisance parameters $(\Lambda, F, \alpha, \xi)$ are drawn from their generative distributions afresh in each replication; only the block-treatment indicator D is held fixed across replications. By the law of total variance, the marginal variance across replications equals $\mathbb{E}_{(\Lambda,F)}[V_t(\Lambda, F, X, D)] + \text{Var}_{(\Lambda,F)}[b_t]$, where $b_t = \mathbb{E}_\varepsilon[\widehat{\text{ATT}}_t - \text{ATT}_t \mid \Lambda, F, X, D]$ is the finite-sample conditional bias. The second term is weakly positive, so the empirical variance upper-bounds $\mathbb{E}_{(\Lambda,F)}[V_t]$; empirical undercoverage therefore implies the procedure under-covers the conditional-variance target of Equation (5) on an average realization, whether the shortfall is the bootstrap's approximation of V_t, finite-sample conditional bias, or both. Reproducibility is ensured by `doRNG` with a single master seed; code and per-cell output are indexed in the `manifest.yaml` at the project root.

A.3.2. Structured-errors results

DGP descriptor		GSC	GSC+split	(i)	(ii)	(iii)
G&A toy, long-range corr, $N_{\text{co}} = 50$	cov	0.960	0.965	0.355	0.365	0.895
	SEr	1.013	1.088	0.248	0.252	0.845
G&A toy, long-range corr, $N_{\text{co}} = 100$	cov	0.955	0.950	0.335	0.360	0.915
	SEr	0.994	1.048	0.216	0.222	0.820
Xu $r = 2$, AR(1) $\rho = 0.8$, $r_{\text{fit}} = 2$	cov	0.920	0.950	0.525	0.550	0.865
	SEr	0.929	0.984	0.384	0.384	0.828
Xu $r = 2$, AR(1) $\rho = 0.8$, $r_{\text{fit}} = 4$	cov	0.985	0.985	0.380	0.405	0.940
	SEr	1.094	1.083	0.218	0.230	0.861

Table A1: Monte Carlo coverage and SE ratio under structured error covariance. "SEr" is the mean bootstrap SE divided by the empirical SD of $\widehat{\text{ATT}}$ across replications; a correctly calibrated procedure has SEr ≈ 1. Monte Carlo SE of coverage ≤ 0.035 across all DGPs. Column heads: GSC = `method="gsynth"` + Xu Alg. 2; GSC+split = GSC with sample splitting on the control pool; (i), (ii), (iii) as in the main text.

A.3.3. i.i.d.-errors results

DGP descriptor		GSC	GSC+split	(i)	(ii)	(iii)
Panel (a): $N_{co} = 50$						
Xu $r = 2$, i.i.d., $r_{\text{fit}} = 2$ (correct)	cov	0.950	0.960	0.975	0.970	0.915
	SEr	1.049	1.128	1.738	1.748	0.927
Xu $r = 2$, i.i.d., $r_{\text{fit}} = 4$ (mild over-spec)	cov	0.900	0.920	0.940	0.955	0.890
	SEr	0.773	0.826	1.501	1.534	0.717
Xu $r = 2$, i.i.d., $r_{\text{fit}} = 6$ (severe over-spec)	cov	0.905	0.915	0.965	0.960	0.900
	SEr	0.787	0.839	1.622	1.700	0.825
Panel (b): $N_{co} = 100$						
Xu $r = 2$, i.i.d., $r_{\text{fit}} = 2$ (correct)	cov	0.955	0.965	0.945	0.945	0.925
	SEr	1.031	1.068	1.068	1.053	0.927
Xu $r = 2$, i.i.d., $r_{\text{fit}} = 4$ (mild over-spec)	cov	0.955	0.955	0.935	0.940	0.900
	SEr	1.013	1.057	1.025	1.020	0.834
Xu $r = 2$, i.i.d., $r_{\text{fit}} = 6$ (severe over-spec)	cov	0.930	0.920	0.860	0.865	0.845
	SEr	0.813	0.849	0.764	0.762	0.703

Table A2: Monte Carlo coverage and SE ratio under i.i.d. errors at the Xu (2017) $r = 2$ factor structure. Panel (a) at $N_{\text{co}} = 50$; Panel (b) at $N_{\text{co}} = 100$. Monte Carlo SE of coverage ≤ 0.035.

Under any error serial correlation, IFE-EM Variants (i) and (ii) drastically undercover, with coverage in $[0.34, 0.55]$ and SE ratios in $[0.22, 0.38]$. The G&A implementation fix

(Variant (ii)) is indistinguishable from the pre-v1.3.1 implementation (Variant (i)) at this scale: at most two to three percentage points of coverage. Variant (iii) (sample splitting on the control pool) provides partial protection but does not close the gap, for the structural reason set out in Appendix A.2.3. The two GSC-family procedures are at or above nominal everywhere under serial correlation. Under i.i.d. errors, IFE-EM Variants (i) and (ii) cover near nominal at correctly specified or mildly over-specified rank, degrading gracefully under severe over-specification; this is why the implementation error was not surfaced by routine i.i.d. validation simulations.

A.4. Additional details of the reanalysis

This section reports additional results for each replicated paper.

A.4.1. Gilens et al. (2021)

This subsection shows the diagnostic plots referenced in §4.1 for two of the five outcomes: the top corporate income tax outcome (the cleanest test of the corporate-stake mechanism in the original paper) and the tort-law outcome (`civil100`), whose apparent significance is a degeneracy of the factor fit rather than a reliable ATT. The treatment-adoption pattern is identical across all five outcomes and is shown once in Figure A1.

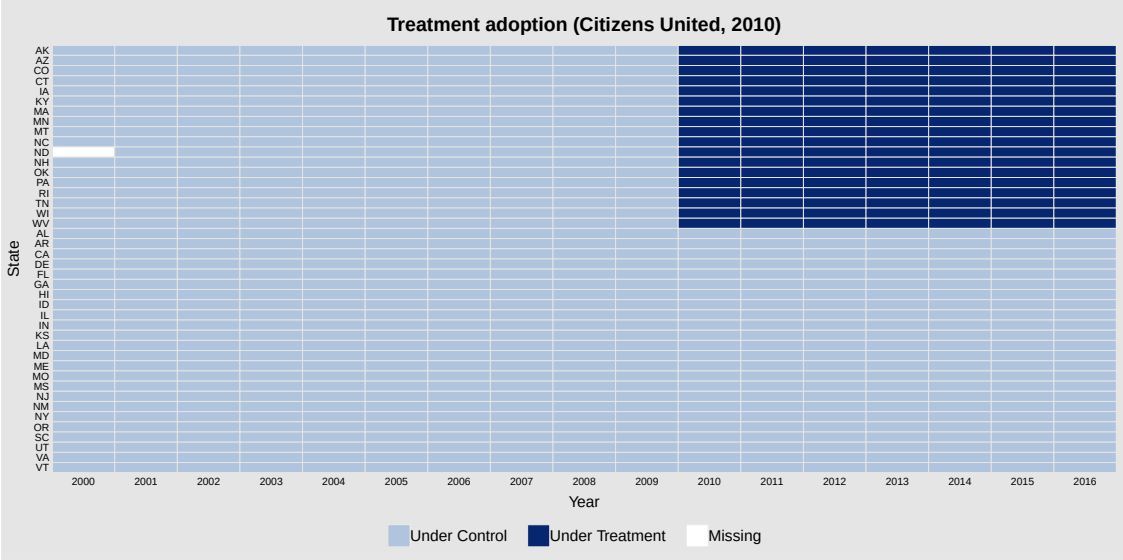


Figure A1: Treatment adoption across the fifty states in Gilens et al. (2021). The pattern applies to all five outcomes.

Top corporate income tax. GSC cross-validation selects $r_{\text{cv}} = 1$, and the resulting average effect is $\widehat{\text{ATT}} \approx -2.9$. The raw outcome has moderate within-state persistence (pooled AR(1) = 0.67); after the GSC fit the residual autocorrelation falls to 0.33—the regime under which the Xu (2017) procedure is designed to operate. Figure A2 shows the per-state bivariate panel; the gap plot separates pre- and post-periods cleanly.

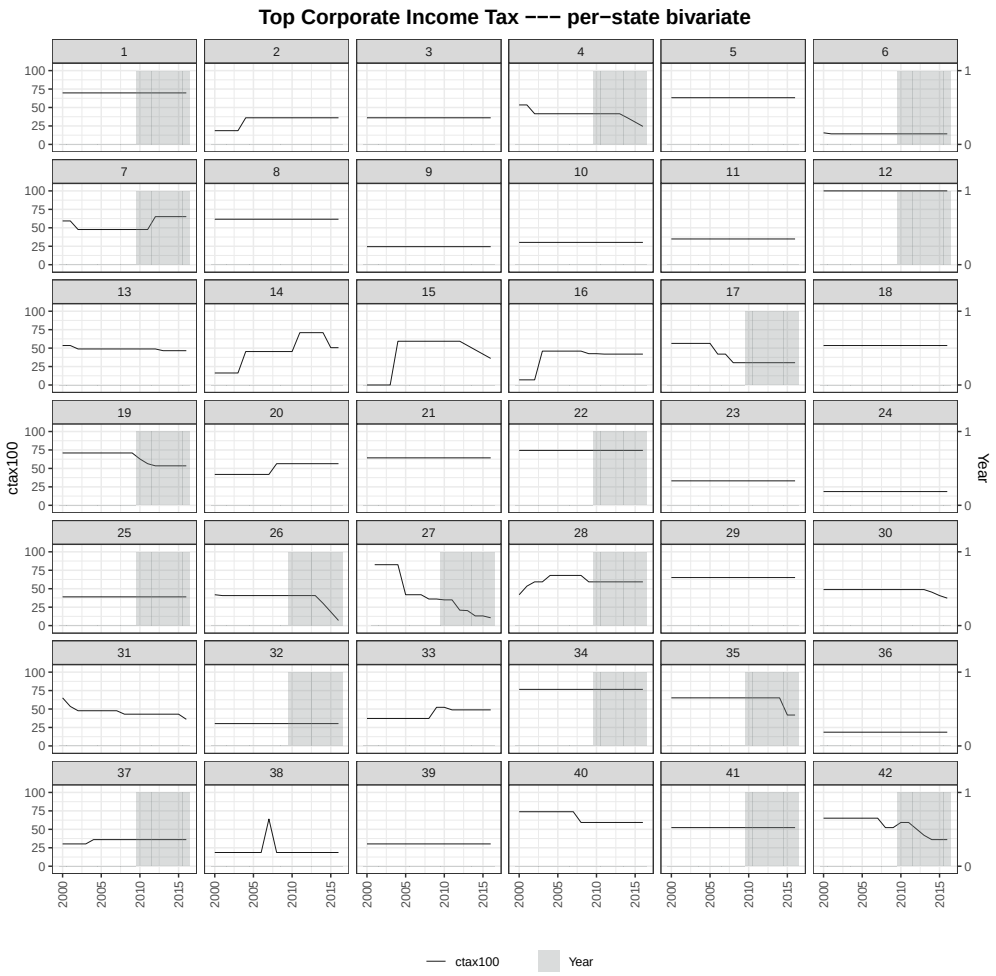
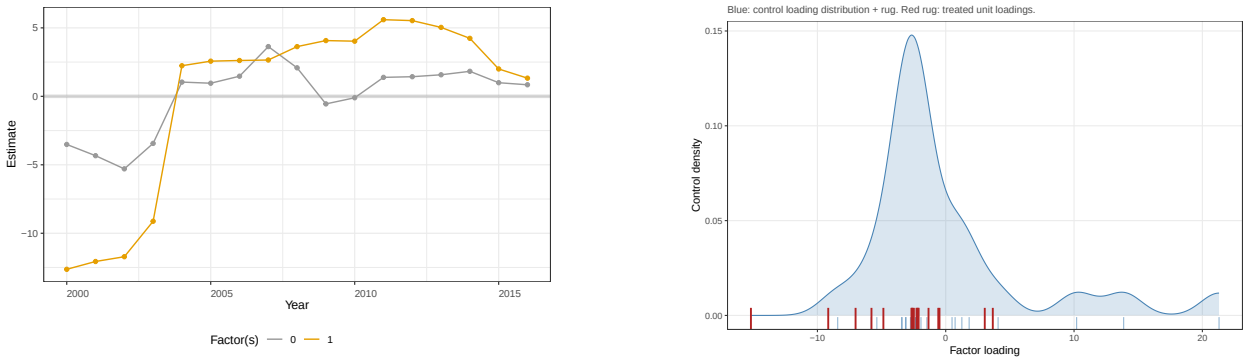


Figure A2: Per-state bivariate panel, top corporate income tax, Gilens et al. (2021).



(a) Estimated factor.

(b) Loading overlap (1D).

Figure A3: Top corporate income tax at CV-selected rank $r = 1$: estimated factor trajectory over time (left) and the loading-overlap diagnostic in the single factor dimension (right). The right panel overlays treated-unit loadings (red rug) on the control-loading distribution (blue density + rug); treated units sitting outside the bulk indicate extrapolation.

Tort law (`civil100`). GSC cross-validation selects $r_{\text{cv}} = 2$, and the resulting average effect is $\widehat{\text{ATT}} \approx -3.2$. The raw outcome has moderate persistence (pooled AR(1) $= 0.65$), essentially identical to the tax outcome; but most states adjust their tort index within a two-year window around 2001–2005 and then remain flat (Figure A4), and the factor model absorbs the step into its first factor. The residual autocorrelation after the fit falls to -0.02—essentially zero, as a degeneracy rather than as a signal of i.i.d. noise.

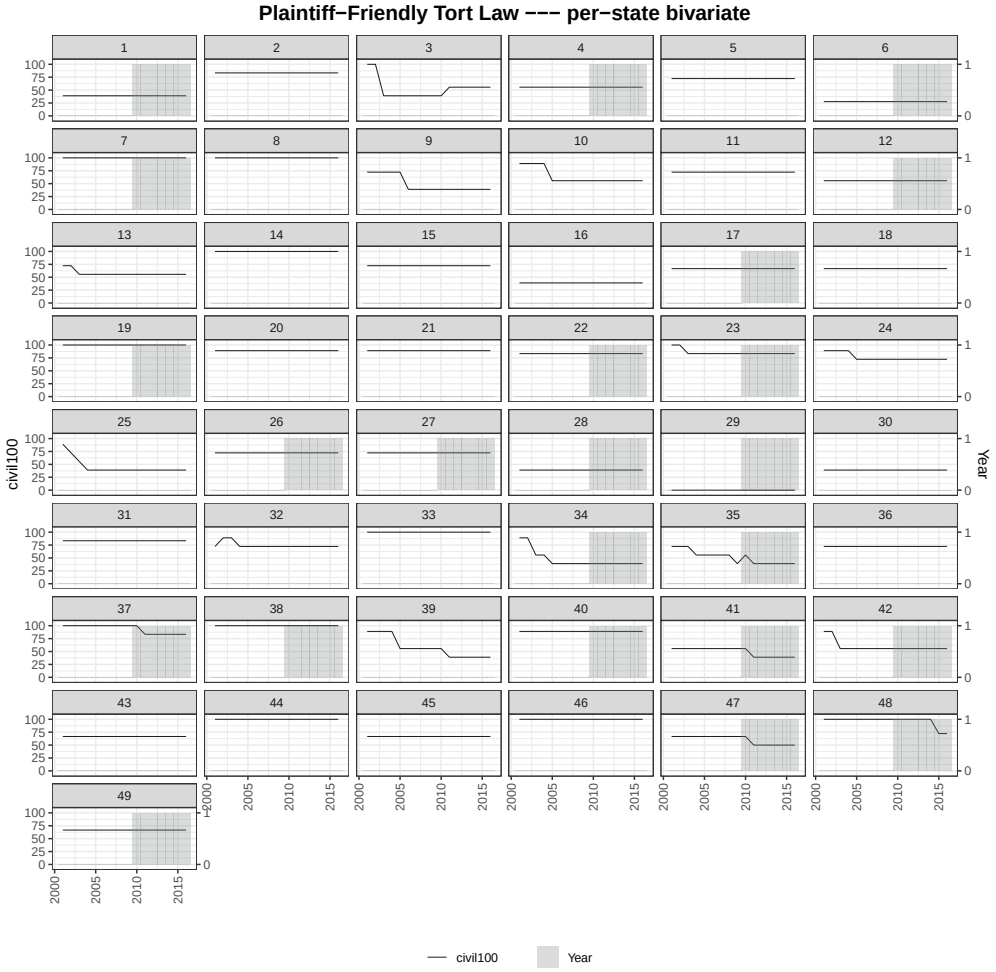
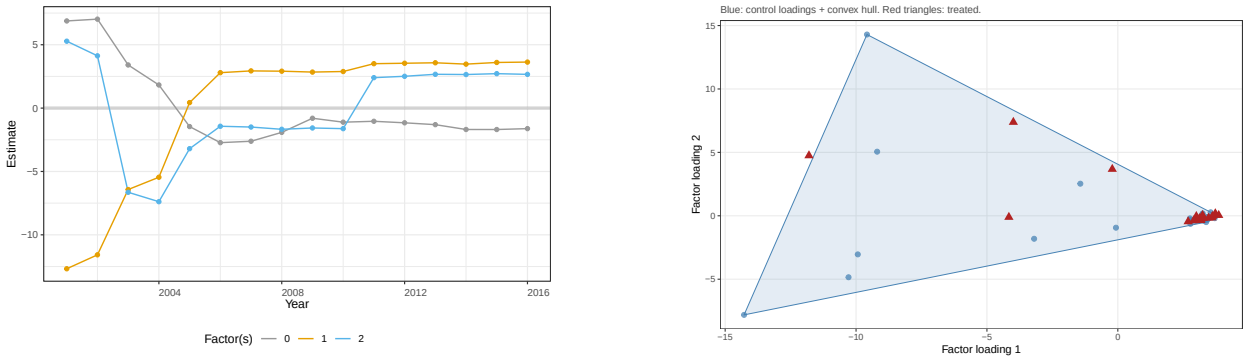


Figure A4: Per-state bivariate panel, tort-law outcome. Most states' outcomes are flat outside a narrow adoption window, the signature of a near-step-function.



(a) Estimated factors.

(b) Loading overlap.

Figure A5: Tort-law outcome at CV-selected rank $r = 2$: estimated factor trajectories over time (left) and the loading-overlap diagnostic in the first two factor dimensions (right). The right panel scatters control-unit loadings in blue with the convex hull shaded, and treated-unit loadings as red triangles. The first factor absorbs the step-function signal in `civil100`.

Forest at the original CV picks (G&A comparison). Figure A6 shows the four-procedure forest at the rank Gilens et al. (2021) originally picked under fect's pre-v2.3.0 block CV defaults—$\{r = 1, 1, 2, 1, 2\}$ for tax, abortion, guns, eminent domain, tort respectively—which is also what G&A reanalyzed against. Under the recommended fix (GSC with the Xu (2017) parametric bootstrap and the corrected unit-level Gaussian approximation, §A.5), tax loses significance (the original headline tax effect does not survive valid inference) and only the tort outcome remains significant; the three originally null outcomes remain null. The picture is consistent with both the preferred-rank verdict in Figure 3 (the rank choice does not drive the verdict on these cells) and with G&A's reanalysis.

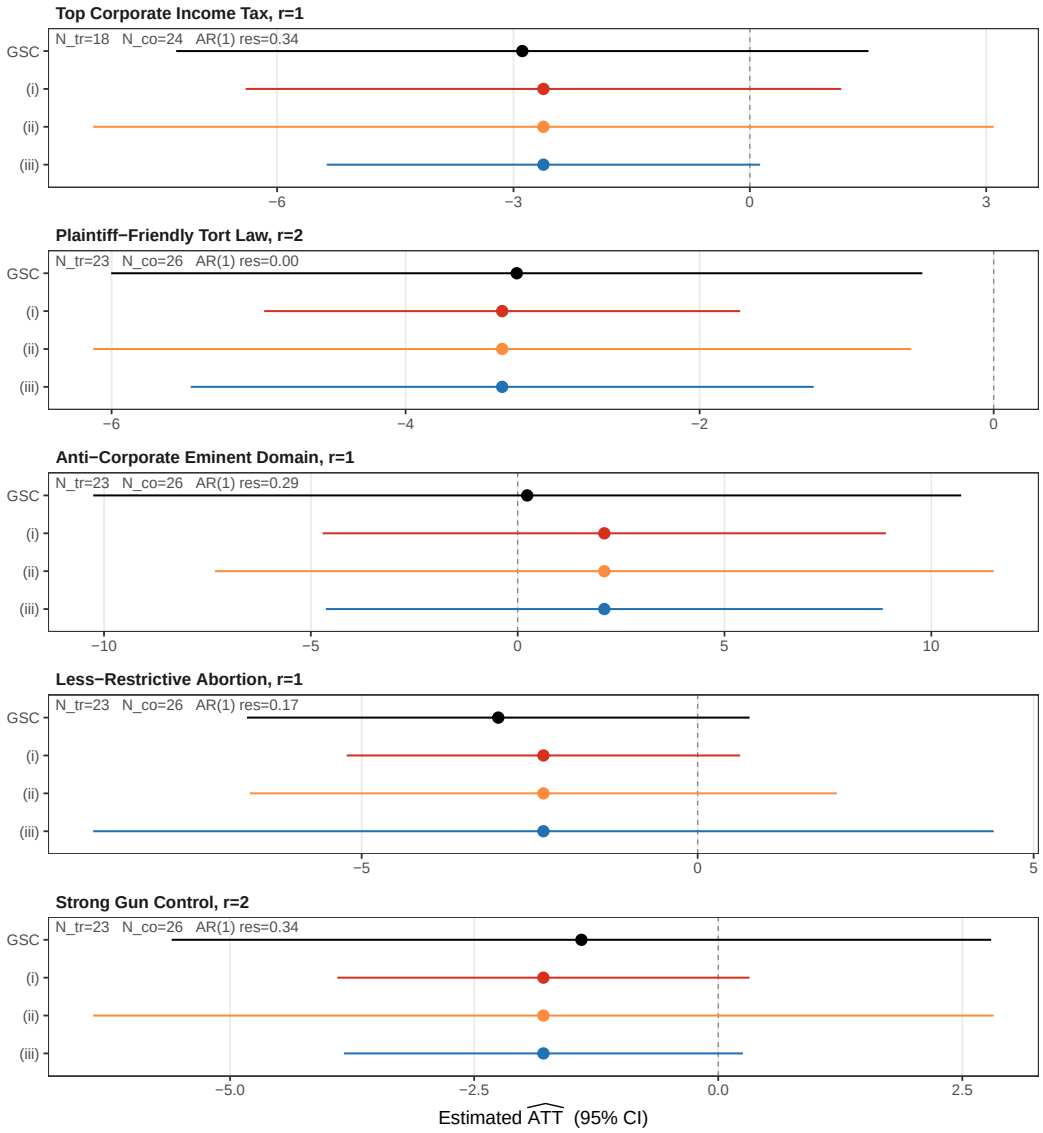


Figure A6: Point estimates and 95% confidence intervals at the original block-CV picks per cell—$\{r = 1, 2, 1, 1, 2\}$ for tax, tort, eminent domain, abortion, gun control respectively—for the five outcomes in Gilens et al. (2021), under GSC with the Xu (2017) parametric bootstrap and IFE-EM Variants (i), (ii), (iii). $B = 200$. Companion to the preferred-rank forest in Figure 3.

A.4.2. Alsaadi (2025)

This subsection shows a rank-sensitivity analysis for the two treatments in the Alsaadi (2025) design. Both share the same outcome (`v2caautmob_osp`, the V-Dem mobilization-autonomy index) and differ in the definition of the minority-regime treatment, which determines the treated sample of countries: Treatment (a) uses `minority` (a single majority excluded), Treatment (b) uses `frac_minority` (other minorities excluded). All fits follow the original paper's specification and use unit fixed effects only (no time fixed effects); $r = 0$ therefore

corresponds to imputation based on unit fixed effects alone. For each treatment, GSC main-effect and placebo-test estimates are reported at four manually-set ranks $r \in \{0, 1, 2, 5\}$; $r = 5$ is the rank CV pegged at and is shown as overfit evidence rather than a recommendation.

Treatment (a): `minority`, $N_{\mathbf{tr}} = 12$. Figures A7 and A8 show treatment adoption and raw outcome trajectories; Figure A9 shows gap (left) and placebo (right) plots at each rank.

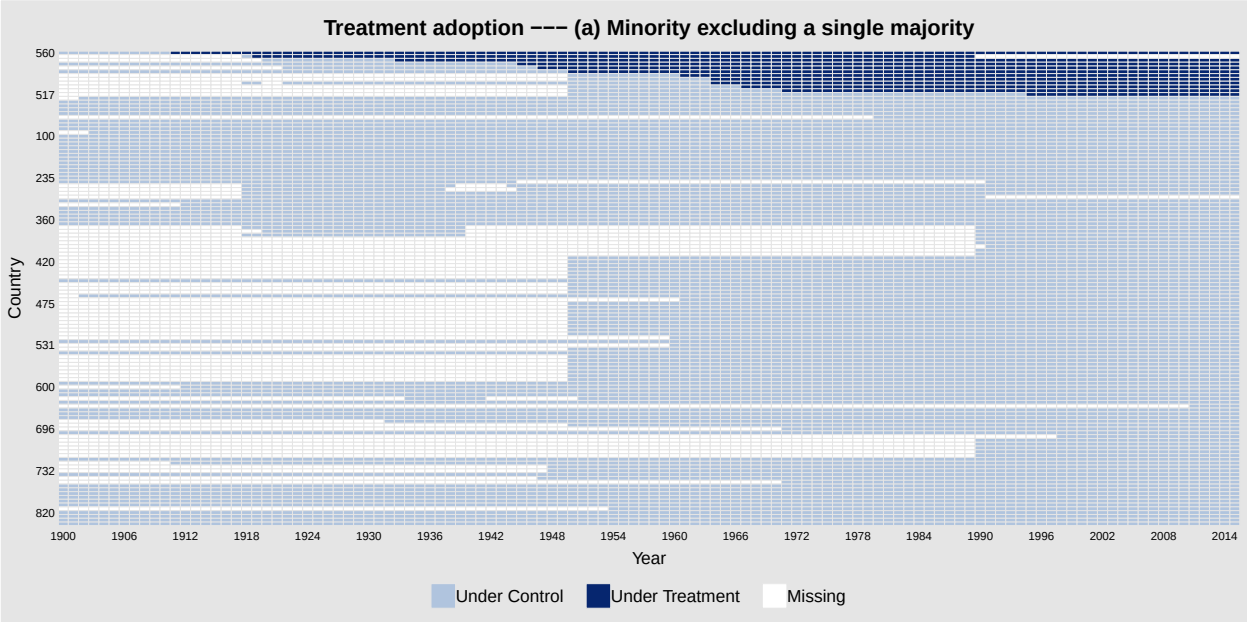


Figure A7: Treatment adoption, Treatment (a) `minority`.

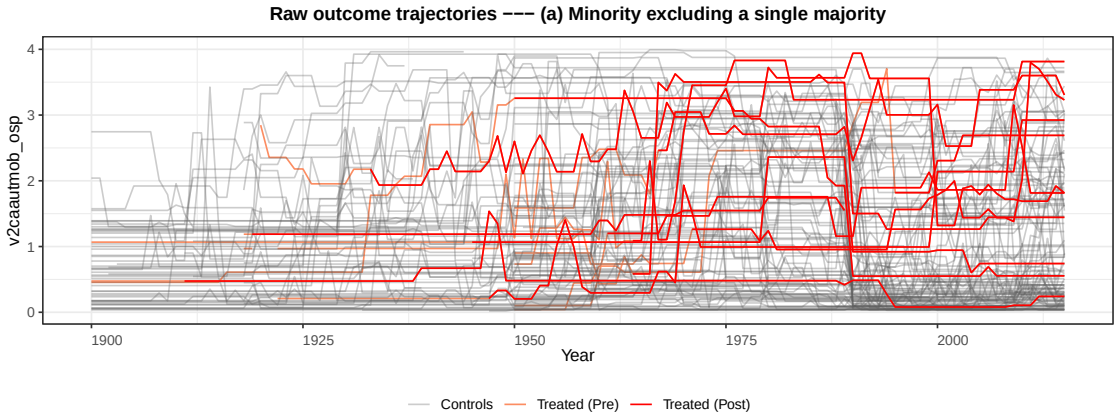
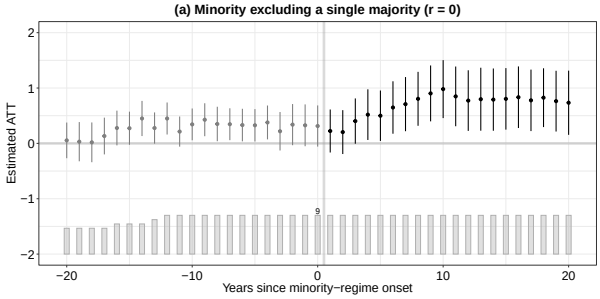
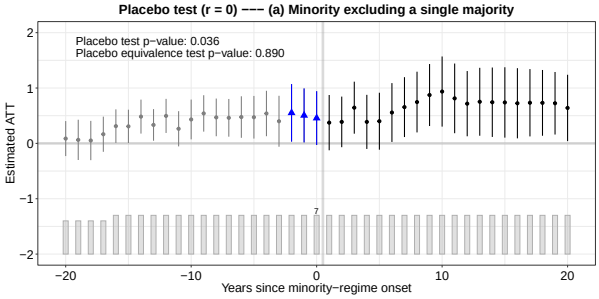


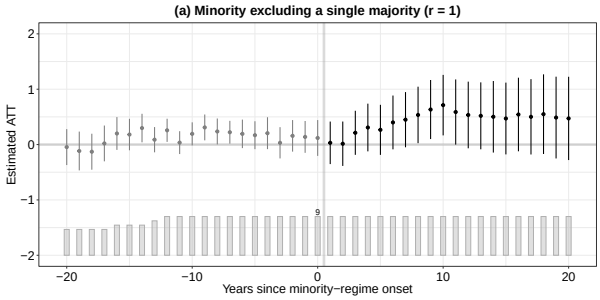
Figure A8: Raw outcome trajectories, Treatment (a) `minority`.



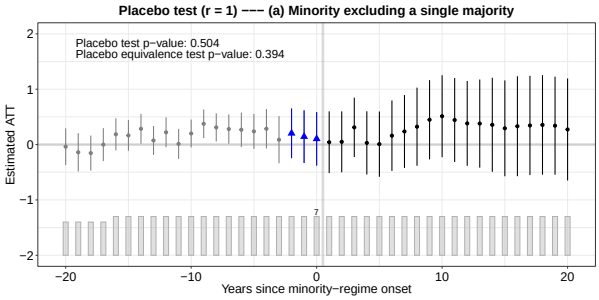
(a) Gap, $r = 0$



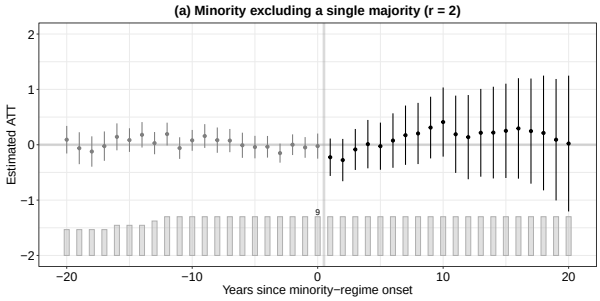
(b) Placebo, $r = 0$



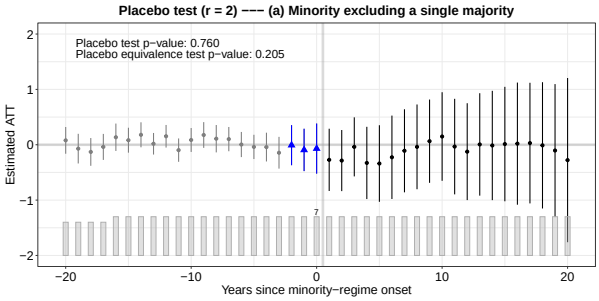
(c) Gap, $r = 1$



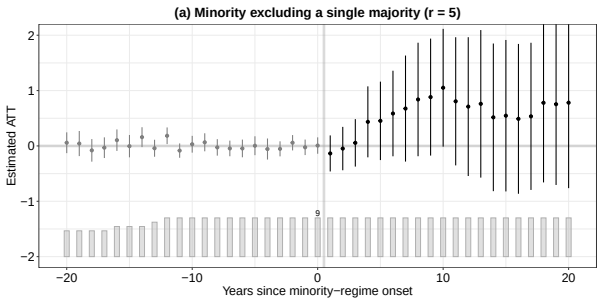
(d) Placebo, $r = 1$



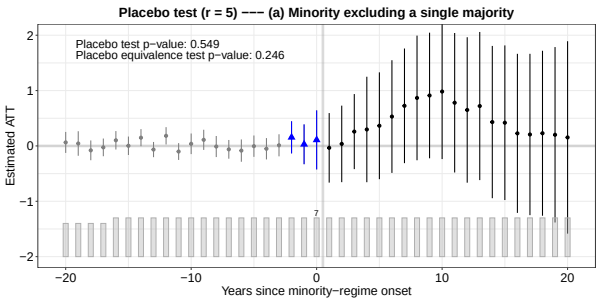
(e) Gap, $r = 2$



(f) Placebo, $r = 2$



(g) Gap, $r = 5$

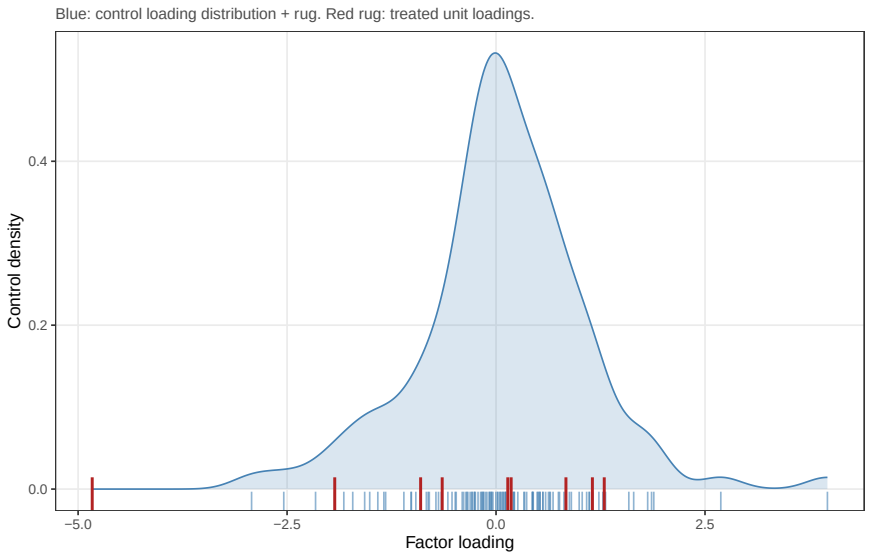


(h) Placebo, $r = 5$

Figure A9: Treatment (a) `minority`: GSC gap (left) and placebo (right) plots at $r \in \{0, 1, 2, 5\}$.

Figure A10 shows the estimated factor and loadings at the preferred rank $r = 1$.

(a) Estimated factor.



(b) Loading overlap (1D).

Figure A10: Treatment (a) `minority` at $r = 1$: estimated factor trajectory over time (top) and the loading-overlap diagnostic in the single factor dimension (bottom). Treated-unit loadings (red rug) are overlaid on the control-loading distribution (blue density + rug); treated units sitting outside the bulk indicate extrapolation.

Treatment (b): `frac_minority`, $N_{\mathbf{tr}} = 24$. Figures A11 and A12 show treatment adoption and raw outcome trajectories; Figure A13 shows gap (left) and placebo (right) plots at each rank.

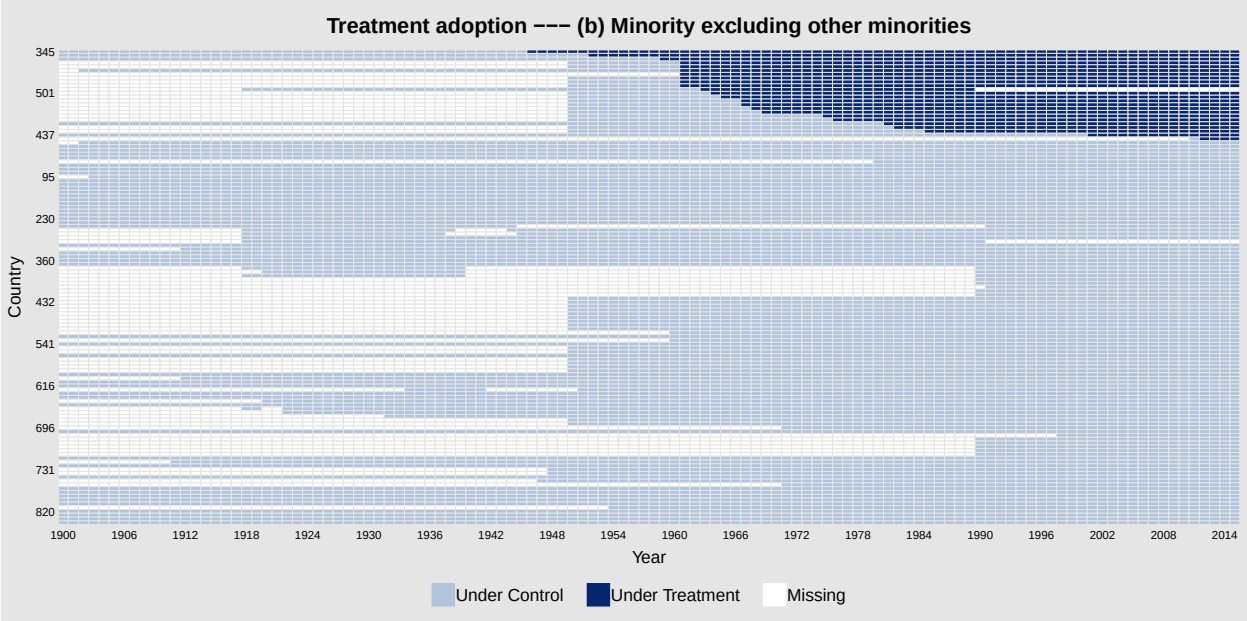


Figure A11: Treatment adoption, Treatment (b) `frac_minority`.

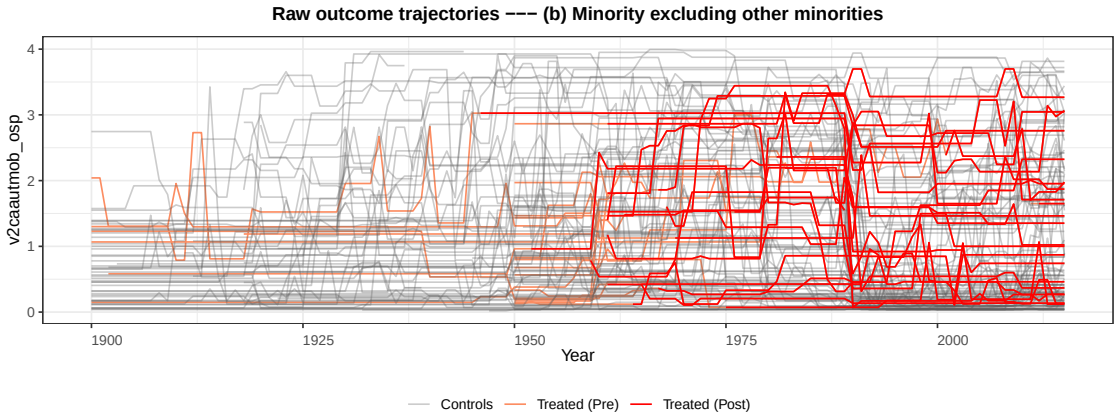
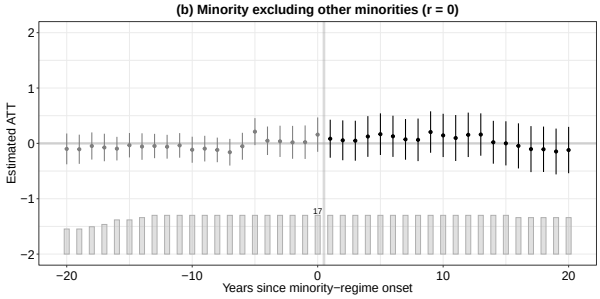
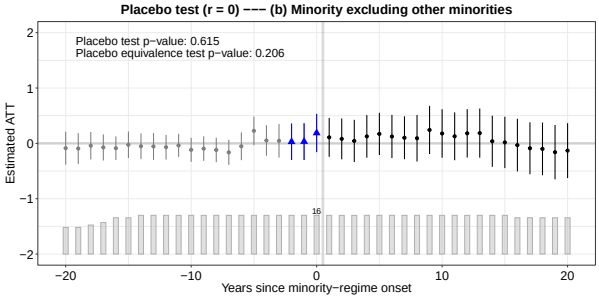


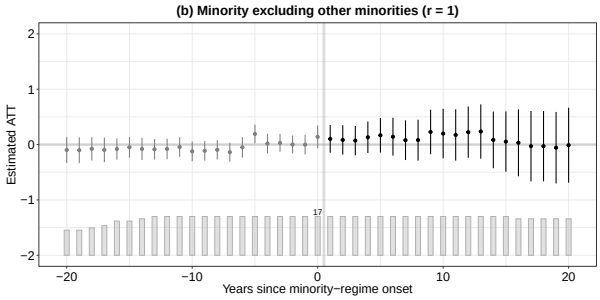
Figure A12: Raw outcome trajectories, Treatment (b) `frac_minority`.



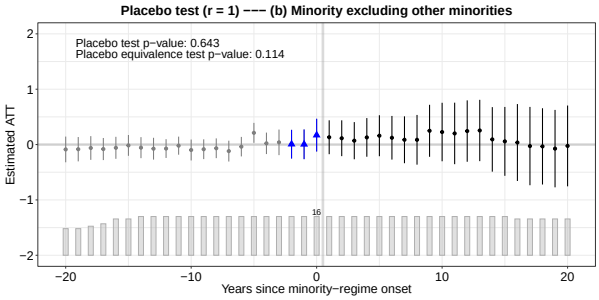
(a) Gap, $r = 0$



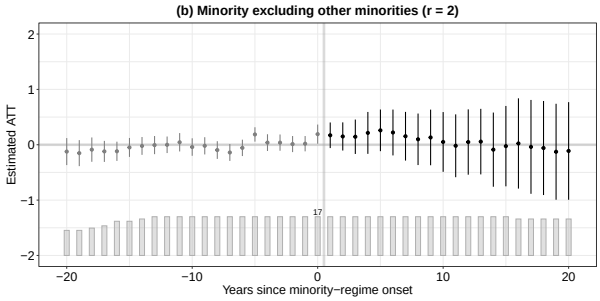
(b) Placebo, $r = 0$



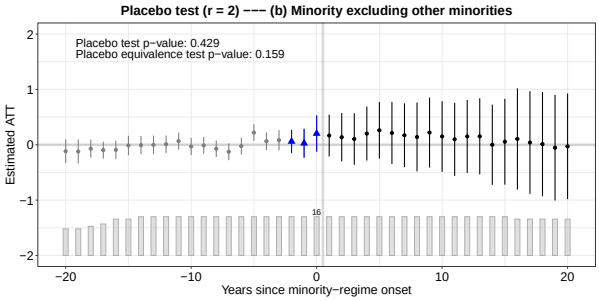
(c) Gap, $r = 1$



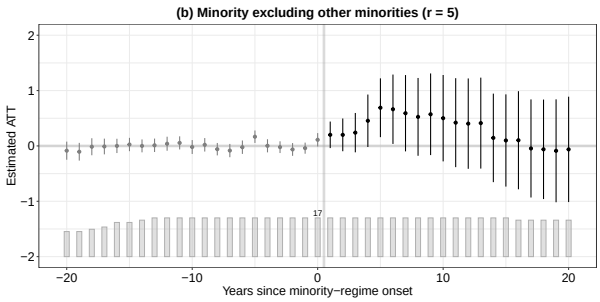
(d) Placebo, $r = 1$



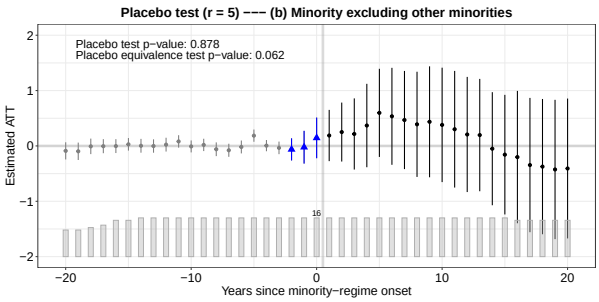
(e) Gap, $r = 2$



(f) Placebo, $r = 2$



(g) Gap, $r = 5$



(h) Placebo, $r = 5$

Figure A13: Treatment (b) `frac_minority`: GSC gap (left) and placebo (right) plots at $r \in \{0, 1, 2, 5\}$.

Findings. For Treatment (a), at $r = 0$ there are clear signs of unobserved confounding: the placebo coefficient is large (+0.50, comparable to the ATT) and marginally rejects at the 5% level ($p \approx 0.04$). These signs disappear once the model includes one or two factors, and at neither $r = 1$ nor $r = 2$ can a significant effect on the outcome be detected. At $r = 5$ the model is clearly overfit. For Treatment (b), the main effect is close to zero at every rank, the placebo test does not reject at every rank, and no factor structure is required ($r = 0$, imputation based on unit fixed effects alone, is sufficient). Table A3 records the exact estimates.

	Main ATT				Placebo effect (periods $-2, -1, 0$)			
r	ATT	SE	95% CI	p	Effect	SE	95% CI	p
Treatment (a): `minority`*, excluding a single majority* ($N_{tr} = 12$)								
0	+0.515	0.247	[+0.032, +0.998]	0.037	+0.504	0.240	[+0.034, +0.975]	0.036
1	+0.345	0.470	[−0.576, +1.267]	0.463	+0.150	0.225	[−0.290, +0.590]	0.504
2	+0.062	0.661	[−1.233, +1.358]	0.925	−0.057	0.186	[−0.421, +0.308]	0.760
5	+0.234	0.768	[−1.271, +1.740]	0.760	+0.098	0.163	[−0.223, +0.418]	0.549
Treatment (b): `frac_minority`*, excluding other minorities* ($N_{tr} = 24$)								
0	+0.054	0.170	[−0.280, +0.388]	0.753	+0.083	0.165	[−0.240, +0.406]	0.615
1	+0.106	0.312	[−0.505, +0.717]	0.735	+0.061	0.130	[−0.195, +0.316]	0.643
2	+0.093	0.390	[−0.672, +0.857]	0.812	+0.096	0.122	[−0.143, +0.335]	0.429
5	+0.084	0.436	[−0.770, +0.937]	0.848	+0.020	0.129	[−0.233, +0.273]	0.878

Table A3: Sensitivity of GSC main ATT and placebo estimates to the factor rank r for the two Alsaadi (2025) treatments. Placebo column uses `placebo.period = c(-2, 0)`; $B = 200$. The $r = 5$ row matches what block CV selects on the paper's grid. Treatment-cell sample sizes shown here ($N_{\text{tr}} = 12, 24$) follow the original paper's convention; the monotonized fits used elsewhere in this note have $N_{\text{tr}} = 9$ and 17 respectively (§4.2).

Forest at $r = 5$ (G&A comparison). Figure A14 shows the four-procedure forest at the rank block CV originally pegged on. The original paper's GSC estimate under (a) was significant; under the recommended fix (GSC with the Xu (2017) parametric bootstrap, §3.2) it is no longer significant. The leave-one-out-corrected IFE-EM bootstrap (Variant (ii) in the figure) reaches the same verdict; §3.1 explains why it is not the procedure we recommend.

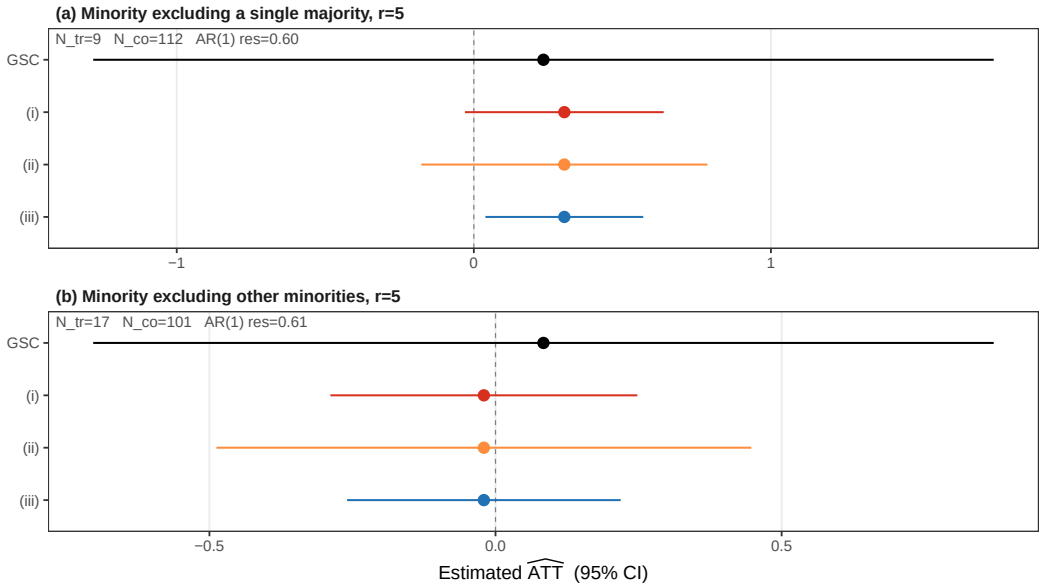


Figure A14: Point estimates and 95% confidence intervals at the original block-CV pick $r = 5$ for the two treatment specifications in Alsaadi (2025), under GSC with the Xu (2017) parametric bootstrap and IFE-EM Variants (i), (ii), (iii). Footers report N_{tr}, N_{co}, and post-fit residual autocorrelation. $B = 200$. Companion to the preferred-rank forest in Figure 5.

A.4.3. Eibl and Hertog (2023)

This subsection reports the full rank-sensitivity grid for the four Eibl and Hertog (2023) outcomes referenced in §4.3: health equity (`v2pehealth_osp`), education equity (`v2peedueq_osp`), primary school enrollment (`priad_ipo`), and secondary school enrollment (`secenrol_combinedplus`). All four share the same oil-rich centre-seeking-subversion treatment; only the outcome variable changes across analyses. Fits use unit fixed effects only, to match the specification of the paper's replication scripts. For each outcome, GSC main-effect and placebo estimates are reported at four manually-set ranks $r \in \{0, 1, 2, 5\}$; $r = 5$ is the rank CV selects under the default `cv.nobs = 3, cv.donut = 1`, and is included as overfit evidence rather than a recommendation.

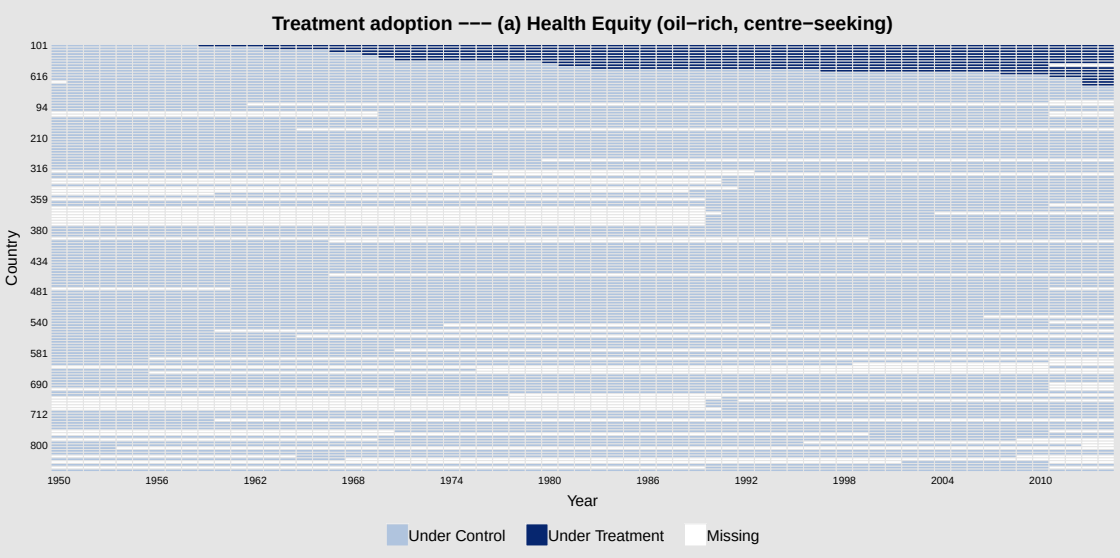
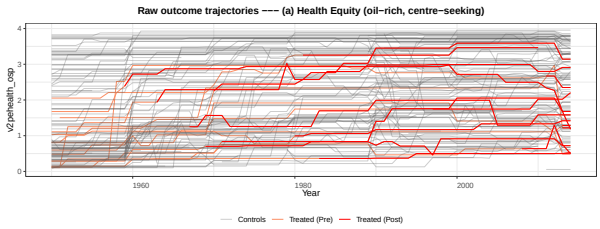
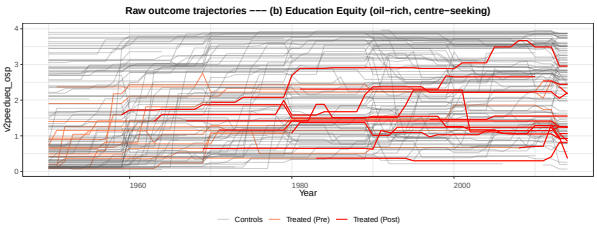


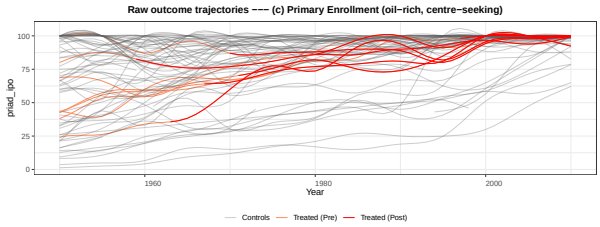
Figure A15: Treatment-adoption pattern in the Eibl and Hertog (2023) specification. The same treated set applies to all four outcomes; only the outcome variable differs.



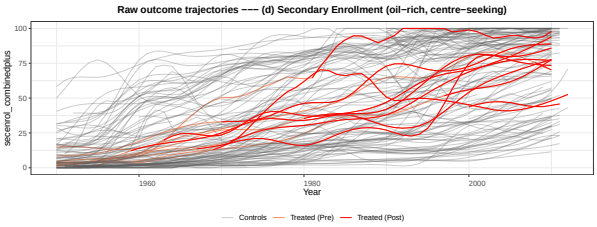
(a) Health equity.



(b) Education equity.



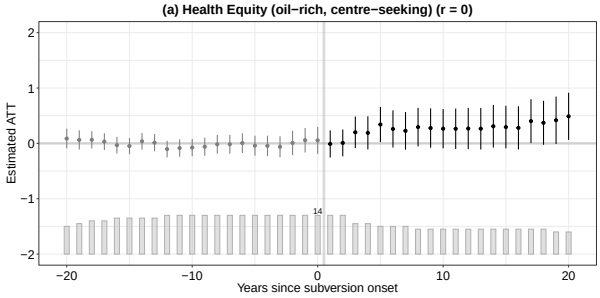
(c) Primary enrollment.



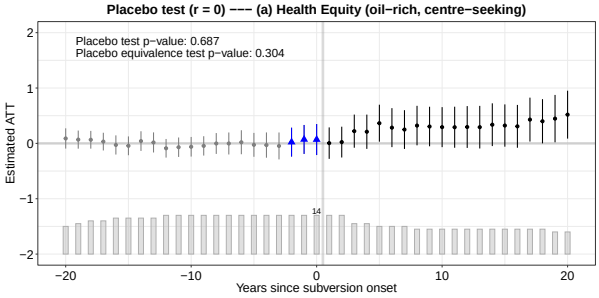
(d) Secondary enrollment.

Figure A16: Raw outcome trajectories for the four Eibl and Hertog (2023) outcomes. Treated-unit trajectories are marked in red; control trajectories in grey. The two welfare indices (health and education equity) are bounded V-Dem ordinal constructs; the two enrollment outcomes are percent-scale.

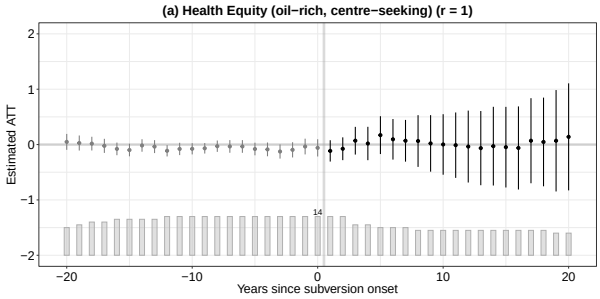
Health equity (`v2pehealth_osp`), $N_{\mathbf{tr}} = 15$. At $r = 0$ (unit fixed effects only) GSC returns $\widehat{\text{ATT}} = +0.40$ ($p = 0.022$) and the placebo test does not reject ($p = 0.69$). Adding one factor degrades the fit: $r = 1$ yields a near-zero ATT, a pattern consistent with the first factor fitting treatment-period structure rather than shared time variation. $r = 2$ and $r = 5$ overshoot in the opposite direction; at $r = 5$ the point estimate sign-flips. Residual autocorrelation at $r = 5$ remains high (≈ 0.57), consistent with factors failing to absorb time structure on this panel.



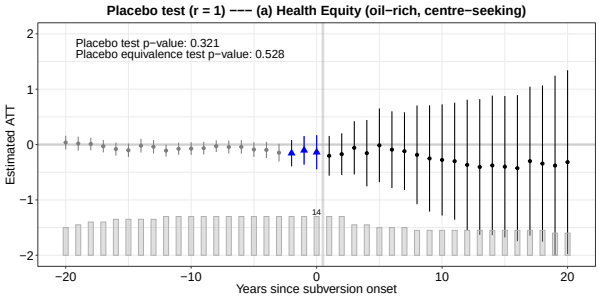
(a) Gap, $r = 0$



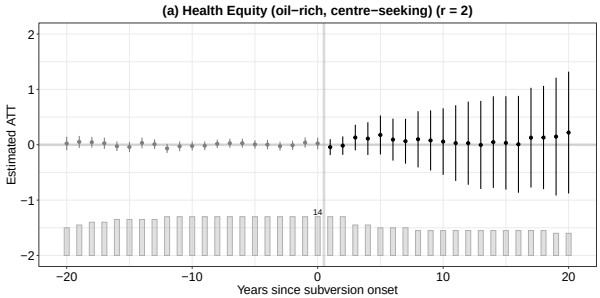
(b) Placebo, $r = 0$



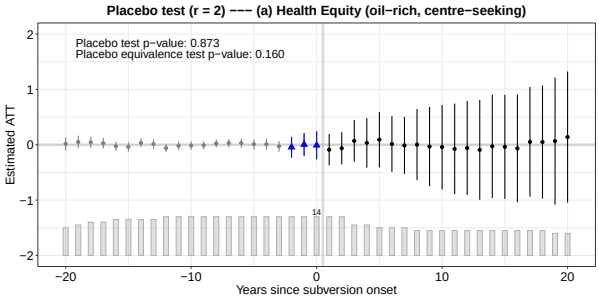
(c) Gap, $r = 1$



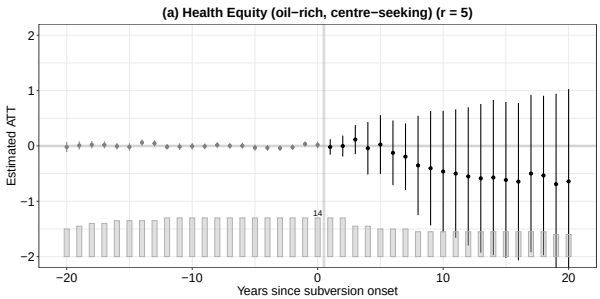
(d) Placebo, $r = 1$



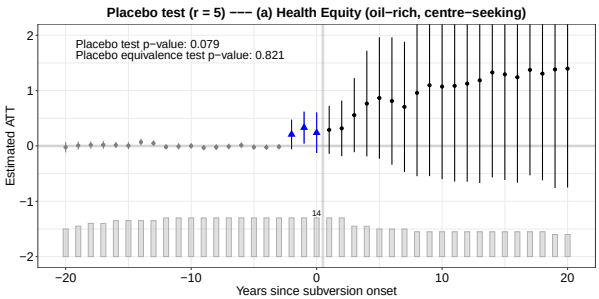
(e) Gap, $r = 2$



(f) Placebo, $r = 2$



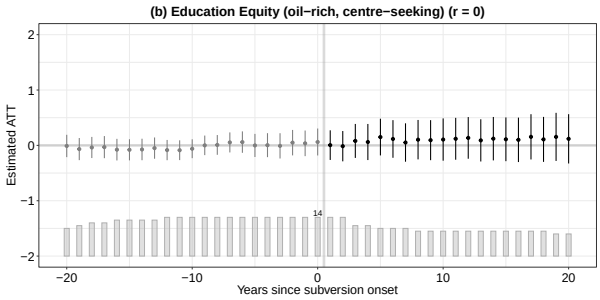
(g) Gap, $r = 5$



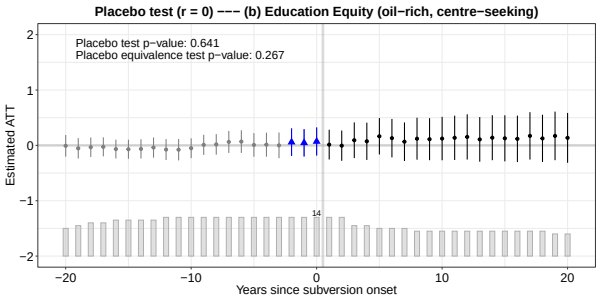
(h) Placebo, $r = 5$

Figure A17: Health equity: GSC gap (left) and placebo (right) plots at $r \in \{0, 1, 2, 5\}$.

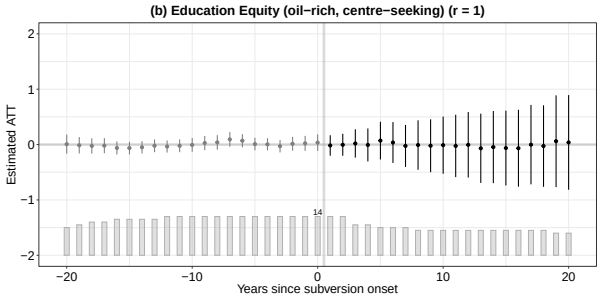
Education equity (`v2peedueq_osp`), $N_{\mathbf{tr}} = 15$. The pattern is the cleanest of the four outcomes. At $r = 0$ the placebo test does not reject ($p = 0.64$) and the main ATT is +0.11, but under the corrected bootstrap the SE is wide enough that the ATT is not statistically significant ($p = 0.55$). Adding one factor produces a null ATT estimate; $r = 2$ and $r = 5$ both sign-flip with large negative point estimates—the classical over-specification signature in the point estimate, though the wide bootstrap SEs at those over-specified ranks leave both statistically null. The residual autocorrelation at $r = 5$ is 0.56: the five-factor fit is not even absorbing the serial structure it is nominally extracting.



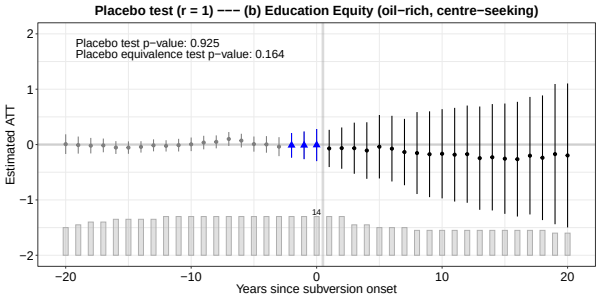
(a) Gap, $r = 0$



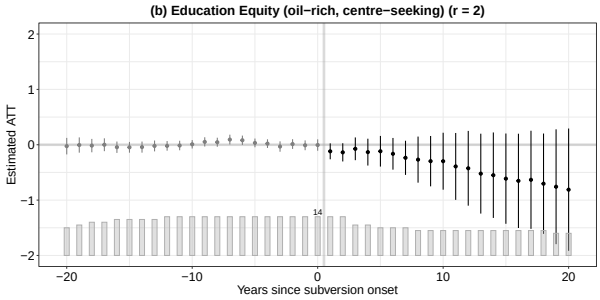
(b) Placebo, $r = 0$



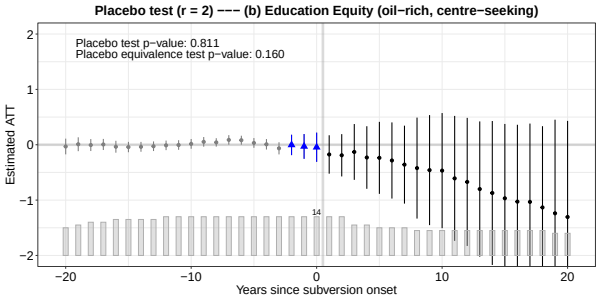
(c) Gap, $r = 1$



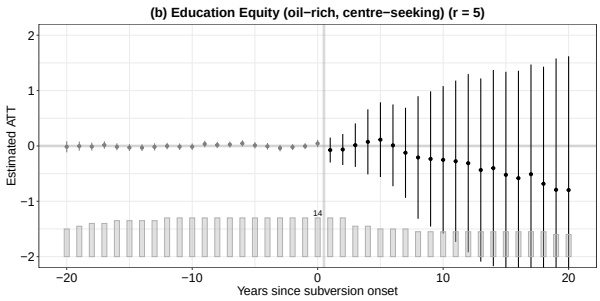
(d) Placebo, $r = 1$



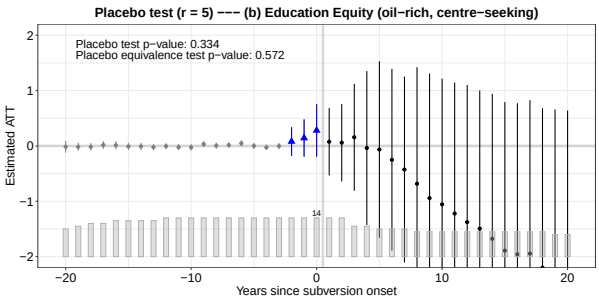
(e) Gap, $r = 2$



(f) Placebo, $r = 2$



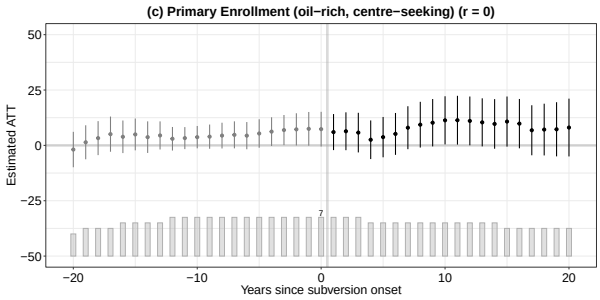
(g) Gap, $r = 5$



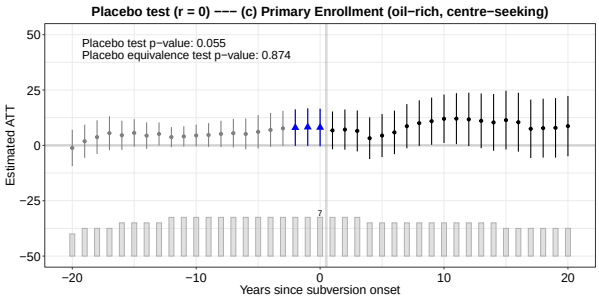
(h) Placebo, $r = 5$

Figure A18: Education equity: GSC gap (left) and placebo (right) plots at $r \in \{0, 1, 2, 5\}$.

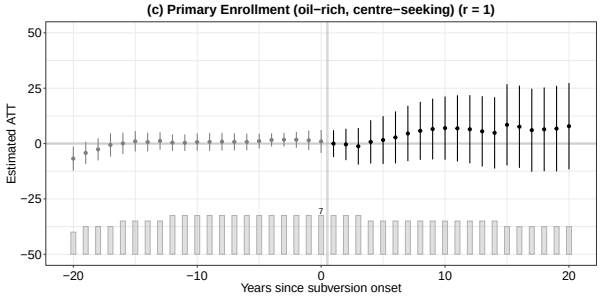
Primary school enrollment (`priad_ipo`), $N_{\mathbf{tr}} = 8$. This is the outcome where a factor structure is genuinely needed. At $r = 0$ the ATT is large and positive (+11.46) but the placebo coefficient is also large (+8.11, comparable in magnitude): without factors the fit is absorbing shared pre-treatment dynamics into apparent treatment effects. At $r = 1$ the placebo coefficient falls to +2.08 ($p = 0.55$), the ATT is +9.46, and the estimate is consistent with the original paper's reported direction (though no longer statistically significant under the corrected bootstrap, $p = 0.32$). Higher ranks over-correct: at $r = 2$ the point estimate sign-flips; at $r = 5$ the main ATT is diffuse with very wide intervals. Residual autocorrelation at $r = 5$ is 0.88, by far the highest of the four outcomes—a diagnostic indication that the factor structure cannot absorb the serial pattern in enrollment regardless of rank.



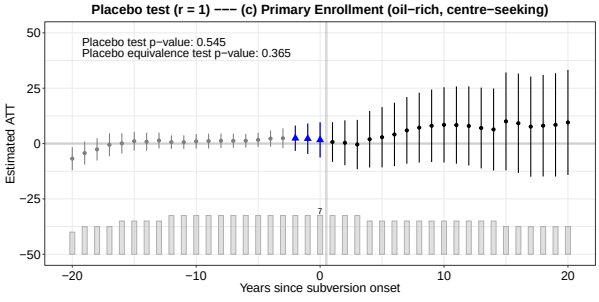
(a) Gap, $r = 0$



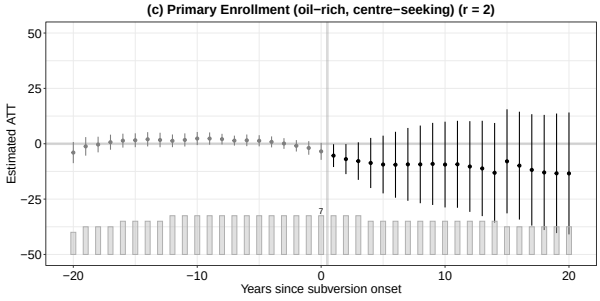
(b) Placebo, $r = 0$



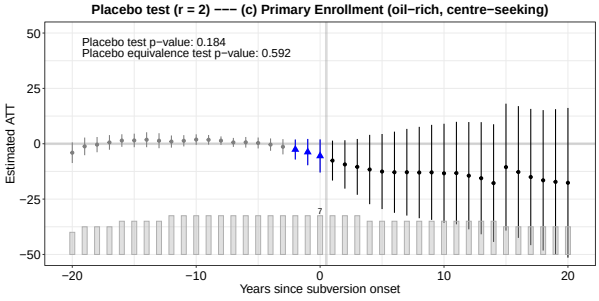
(c) Gap, $r = 1$



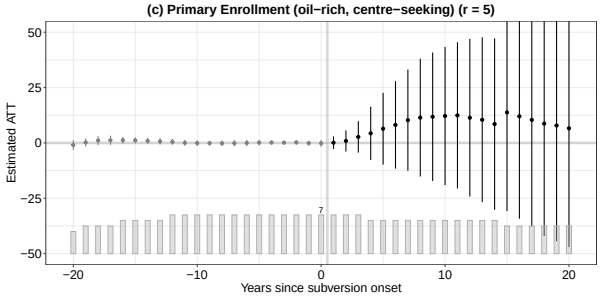
(d) Placebo, $r = 1$



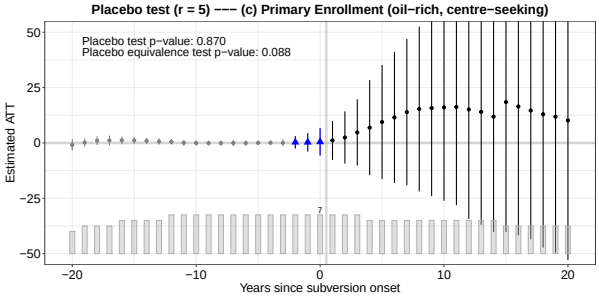
(e) Gap, $r = 2$



(f) Placebo, $r = 2$



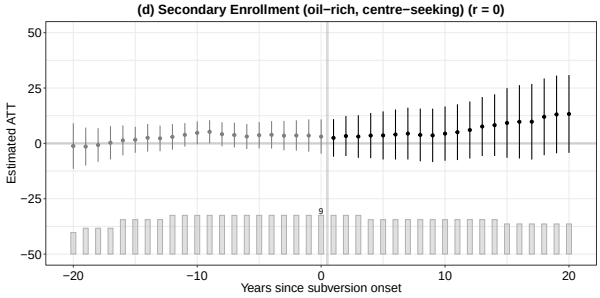
(g) Gap, $r = 5$



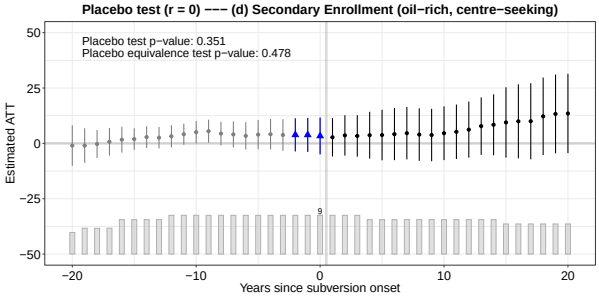
(h) Placebo, $r = 5$

Figure A19: Primary school enrollment: GSC gap (left) and placebo (right) plots at $r \in \{0, 1, 2, 5\}$.

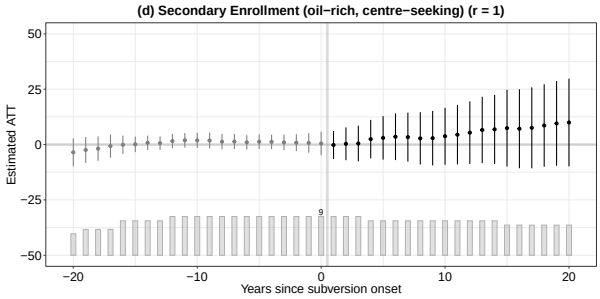
Secondary school enrollment (`secenrol_combinedplus`), $N_{\mathbf{tr}} = 11$. The placebo test does not reject at any rank for this outcome (placebo p-values 0.35 to 0.89), and the main ATT is positive at every rank (+8.59, +6.88, +9.42, +6.39 at $r = 0, 1, 2, 5$). Under the corrected bootstrap, however, the SEs are wide enough that the ATT is statistically null at every rank. At $r = 0$ the placebo coefficient is largest (+3.67); at $r = 1$ it falls to +1.24 and the ATT is +6.88, so $r = 1$ is the preferred specification. Residual autocorrelation at $r = 5$ is 0.81, again very high—factors are not absorbing the time structure here either.



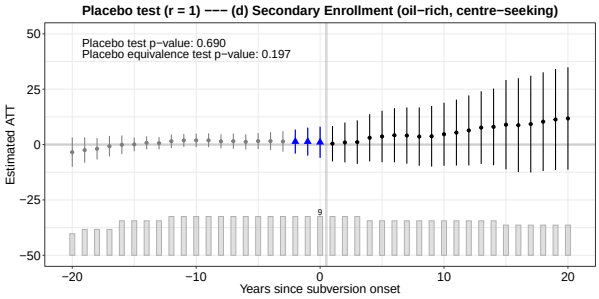
(a) Gap, $r = 0$



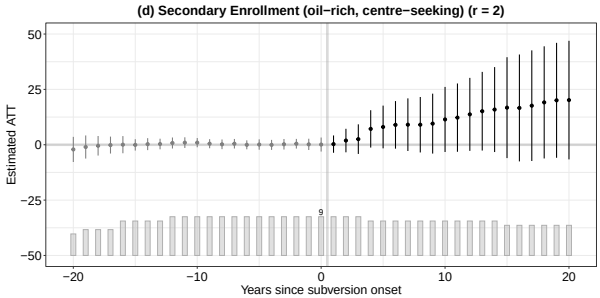
(b) Placebo, $r = 0$



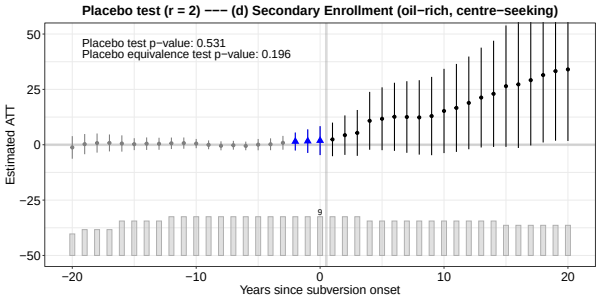
(c) Gap, $r = 1$



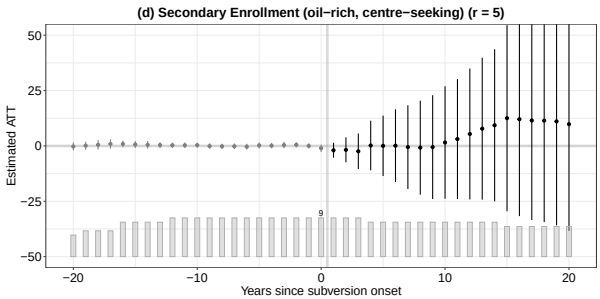
(d) Placebo, $r = 1$



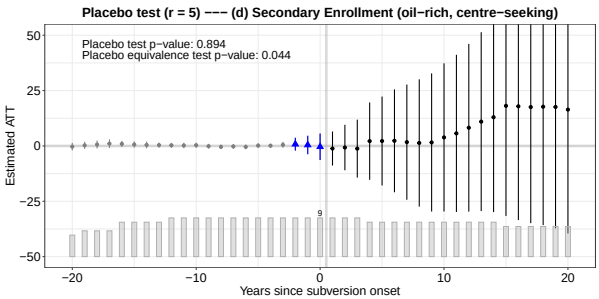
(e) Gap, $r = 2$



(f) Placebo, $r = 2$



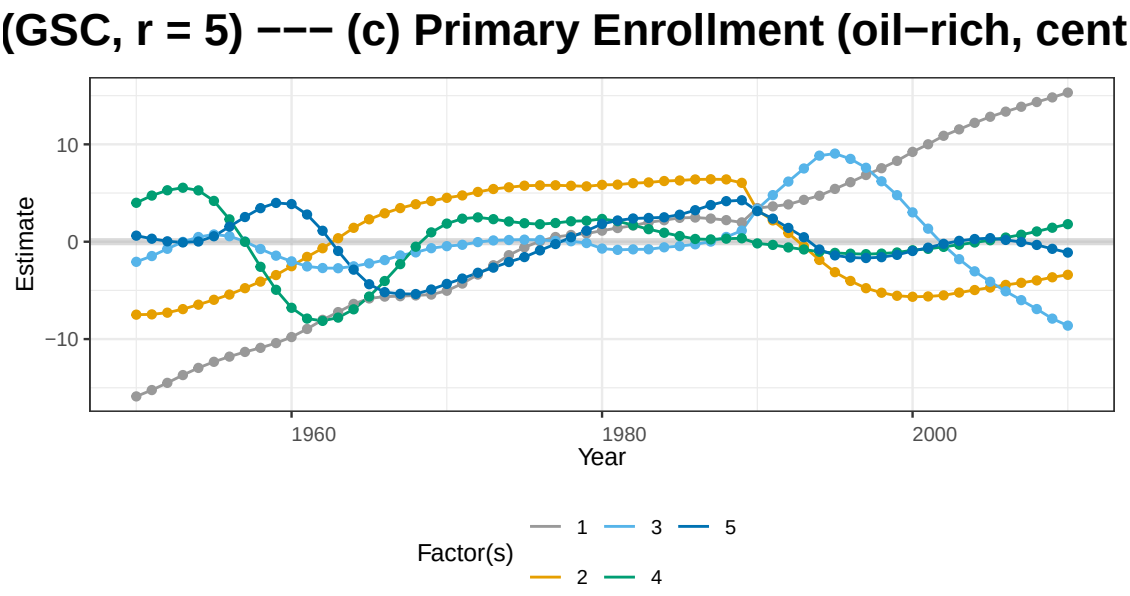
(g) Gap, $r = 5$



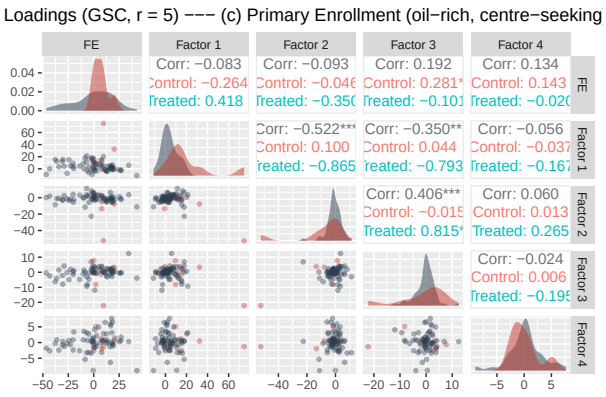
(h) Placebo, $r = 5$

Figure A20: Secondary school enrollment: GSC gap (left) and placebo (right) plots at $r \in \{0, 1, 2, 5\}$.

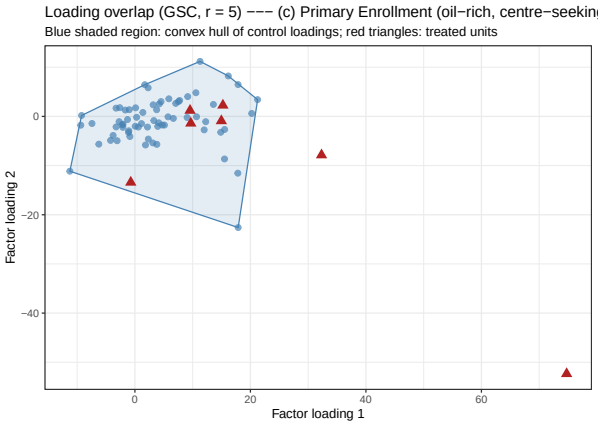
Overfitting at the CV-selected rank: visual diagnostics. Figure A21 shows the estimated factors, loadings, and first-two-factor loading hull for primary enrollment at the CV-selected rank $r = 5$. The five factor trajectories are visually indistinguishable from noise; the loadings are diffuse; the treated units' loadings fall well outside the convex hull of control-unit loadings, indicating that the GSC imputation is extrapolating rather than interpolating. The other three outcomes share the same qualitative pattern at $r = 5$.



(a) Estimated factors at $r = 5$.



(b) Factor loadings at $r = 5$.



(c) First-two-factor loading hull.

Figure A21: Primary school enrollment at the CV-selected rank $r = 5$: estimated factors (top), full factor loadings (bottom left), and first-two-factor loading hull (bottom right). Red points are treated units; grey points are controls; the shaded region is the convex hull of control loadings. The factors look noise-like, loadings are diffuse, and treated units fall outside the control hull—all three are signatures of overfit.

	Main ATT				Placebo effect (periods $-2, -1, 0$)			
r	ATT	SE	95% CI	p	Effect	SE	95% CI	p
Health equity (`v2pehealth_osp`), $N_{tr} = 15$								
0	+0.396	0.173	[+0.058, +0.735]	0.022	+0.054	0.135	[−0.210, +0.318]	0.687
1	+0.000	0.534	[−1.047, +1.048]	0.999	−0.133	0.134	[−0.395, +0.129]	0.321
2	+0.319	0.581	[−0.819, +1.458]	0.583	−0.017	0.107	[−0.226, +0.192]	0.873
5	−0.628	0.693	[−1.987, +0.731]	0.365	+0.259	0.148	[−0.030, +0.548]	0.079
Education equity (`v2peedueq_osp`), $N_{tr} = 15$								
0	+0.112	0.189	[−0.259, +0.483]	0.554	+0.058	0.125	[−0.187, +0.303]	0.641
1	−0.011	0.476	[−0.945, +0.922]	0.981	−0.012	0.127	[−0.261, +0.237]	0.925
2	−0.828	0.626	[−2.055, +0.399]	0.186	−0.026	0.110	[−0.243, +0.190]	0.811
5	−1.283	1.126	[−3.489, +0.923]	0.254	+0.167	0.173	[−0.172, +0.507]	0.334
Primary school enrollment (`priad_ipo`), $N_{tr} = 8$								
0	+11.458	5.748	[+0.191, +22.725]	0.046	+8.107	4.226	[−0.175, +16.389]	0.055
1	+9.459	9.594	[−9.345, +28.263]	0.324	+2.079	3.439	[−4.660, +8.819]	0.545
2	−13.955	15.758	[−44.839, +16.930]	0.376	−3.959	2.980	[−9.801, +1.882]	0.184
5	−10.934	20.791	[−51.684, +29.816]	0.599	+0.353	2.156	[−3.873, +4.578]	0.870
Secondary school enrollment (`secenrol_combinedplus`), $N_{tr} = 11$								
0	+8.588	7.009	[−5.150, +22.326]	0.220	+3.673	3.936	[−4.041, +11.387]	0.351
1	+6.877	8.443	[−9.671, +23.424]	0.415	+1.240	3.113	[−4.862, +7.342]	0.690
2	+9.422	12.885	[−15.831, +34.675]	0.465	+1.646	2.626	[−3.501, +6.794]	0.531
5	+6.386	19.399	[−31.635, +44.407]	0.742	+0.283	2.123	[−3.879, +4.444]	0.894

Table A4: Sensitivity of GSC main ATT and placebo estimates to the factor rank r for the four Eibl and Hertog (2023) outcomes. Placebo column uses `placebo.period = c(-2, 0)`; $B = 200$. The $r = 5$ row matches what block CV selects under `cv.nobs = 3, cv.donut = 1` for all four outcomes.

Forest at $r = 5$ (G&A comparison). Figure A22 shows the four-procedure forest at the rank block CV originally pegs at on every cell. Under the recommended fix—GSC with the Xu (2017) parametric bootstrap—all four within-cell findings are non-significant at $r = 5$.[A2] The pre-v1.3.1 IFE-EM bootstrap (Variant (i)) and its leave-one-out correction (Variant (ii)) retain narrow SEs and remain significant on all four outcomes; Variant (ii) does not reverse the headline at $r = 5$, illustrating the limitation discussed in §3.1.

[A2]The health-equity point estimate sign-flips to negative at $r = 5$. This is an overlap pathology—treated-unit loadings fall outside the convex hull of control loadings under the over-specified factor model—of the kind the bounded-loading algorithm in the latest `fect` (§5) addresses directly. For secondary enrollment at $r = 5$, the in-house Variant (ii) helper falls back to in-sample residuals for cells where the leave-one-out refit cannot produce a clean OOS prediction error on this heavily unbalanced panel (∼27% residual missingness); the resulting SE is therefore mildly downward-biased relative to a clean variant (ii) estimate.

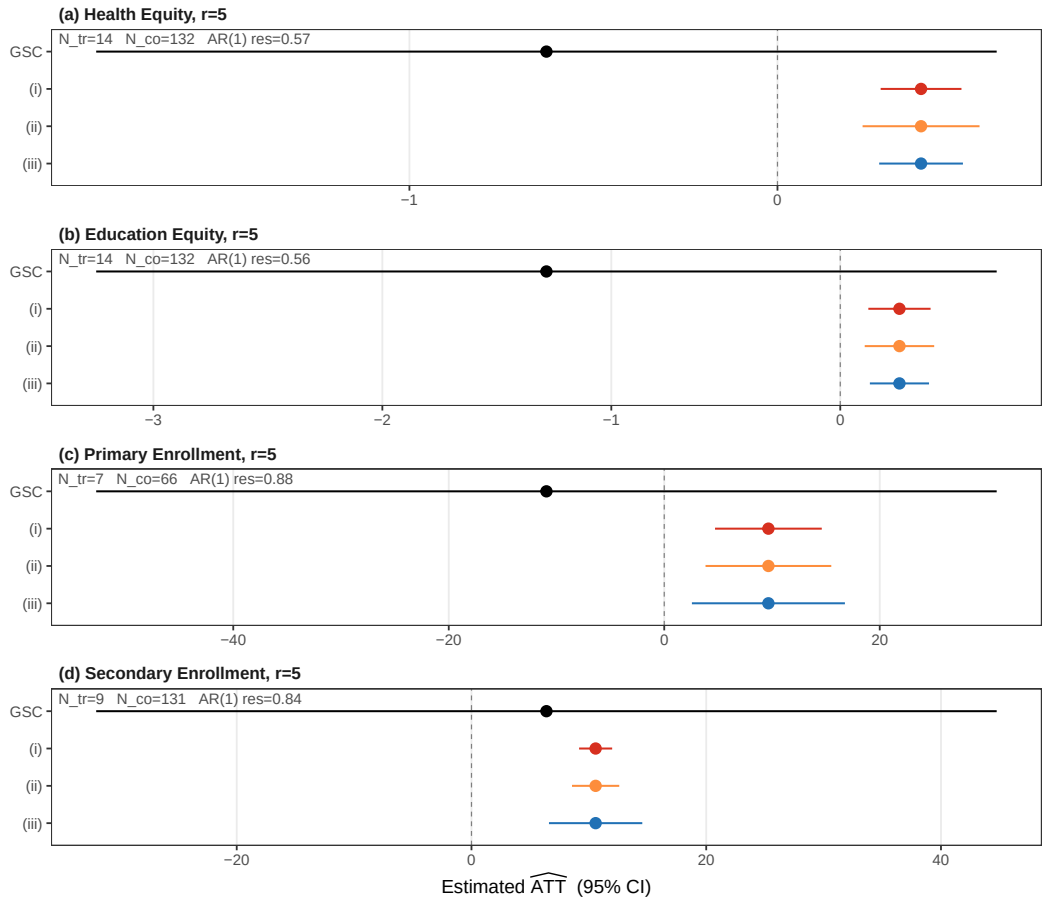


Figure A22: Point estimates and 95% confidence intervals at the original block-CV pick $r = 5$ for the four Eibl and Hertog (2023) outcomes, under GSC with the Xu (2017) parametric bootstrap and IFE-EM Variants (i), (ii), (iii). $B = 200$. Companion to the preferred-rank forest in Figure 8.

Finding. Under the preferred-rank rule used throughout this section, the four Eibl and Hertog (2023) outcomes split: only health equity recovers a significant positive ATT (+0.40 at $r = 0$); education equity, primary enrollment, and secondary enrollment yield point estimates in the original direction but, under the corrected AR-preserving bootstrap, with confidence intervals that cross zero (see the updated forest plot, Figure 8). At the CV-selected rank $r = 5$ all four outcomes are non-significant under GSC (Figure A22); two also sign-flip. The four-to-five-factor gap between the preferred and CV-selected ranks is the largest among the three applications in this section; combined with residual autocorrelation of 0.56 to 0.88 after the $r = 5$ fit, it is the clearest single illustration of the compound mechanism that motivates the note. Under the rank chosen by CV with the pre-v1.3.1 parametric bootstrap, the original within-cell pattern of four positive significant findings does not survive valid inference—and this is the reversal that G&A report.

A.5. Implementation notes for `fect`

This subsection records the rationale for the package-side changes since the v1.3.1 refactor. Each maps to an analytical finding earlier in the note.

API refactor (v2.0.0). Pre-v1.3.1, `gsynth` exposed the GSC and IFE-EM estimators implicitly: a single `method` argument selected one or the other, and the variance procedure was tied to that choice in code paths that were not documented separately. The `fect` v2.0.0 refactor unified the two estimators under a single API and made the structural distinction an explicit parameter: `time.component.from="nevertreated"` restricts factor estimation to control units (the GSC objective of §3.2), while `time.component.from="notyettreated"` extends it to treated units' pre-treatment periods through the IFE-EM imputation step (§3.1). The variance-estimation procedure is now an independent parameter, `vartype`, with values `"parametric"` (the Xu (2017) LOO-based bootstrap), `"bootstrap"` (nonparametric), and `"jackknife"`. Separating the two parameters made explicit a constraint that was implicit in the older code: the parametric SE was designed around the GSC objective, and the validity argument for it (§3.2) relies on insulation properties that do not hold under IFE-EM. The hard gate described next sits at exactly this parameter intersection.

Hard gate on IFE-EM + parametric (v2.2.0). Requesting `vartype="parametric"` together with `time.component.from="notyettreated"` triggers an error on construction, with an explicit message pointing the user to `vartype="bootstrap"` or `vartype="jackknife"`. A hard error rather than a warning was deliberate: a warning is dismissed without action, while an error forces the analyst to make a conscious choice between switching SE procedure, switching estimator to GSC, or accepting that the parametric SE under IFE-EM has no theoretical backing for their setting (§3.1). An auto-fallback to nonparametric SE was also considered and rejected for the same reason: silently changing the SE procedure would mask the issue from analysts who deliberately chose IFE-EM and may not notice the change.

Rolling-window CV default (v2.3.0). The default cross-validation switched from block CV to rolling-window CV (`cv.method="rolling"`); the previous block design remains accessible via `cv.method="block"` for replication. Motivation: §3.4—block CV's cross-fold leakage causes rank inflation under serial correlation. The rolling design cuts the training panel forward in time at a random anchor with a short buffer of cells immediately before the anchor dropped from training, breaking the leakage and substantially mitigating rank inflation under AR(1) errors. Long-range correlation requires the modeling-stage approach of §5 rather than further CV tuning.

Bounded-loading algorithm (v2.3.0). An optional constrained variant of GSC's loading-projection step is now available, restricting each treated unit's projected loading to lie within the convex hull of control-unit loadings. Motivation: §5 (Overlap and the simplex constraint)—when treated loadings sit outside the control hull, the factor-model counterfactual is an extrapolation, and over-specified ranks make this routine (e.g. Eibl and Hertog (2023) health equity at $r = 5$, Appendix A.4.3). The bounded variant is not on by default, since it changes the estimator's objective rather than its diagnostics.

Loading-overlap diagnostic plot (v2.3.0). The `plot()` method on a fitted `fect` object now offers a loading-overlap diagnostic: a scatter of estimated treated- and control-unit loadings in the first two factor dimensions, with the convex hull of control loadings shaded. Motivation: §4—across the three reanalyses, factor-loading overlap was the most discriminating diagnostic between settings where the GSC counterfactual lies inside the data's support and settings where it requires extrapolation. The plot is intended as an at-fit check; the bounded-loading algorithm above is the corresponding remedy when overlap is poor.

Unit-level Gaussian approximation for the GSC parametric bootstrap on unbalanced panels (v2.3.0). The GSC parametric bootstrap has always drawn residual perturbations at the unit level (whole T-vectors at a time, preserving within-unit serial structure). On balanced panels this works as documented. On unbalanced panels with NA cells in the post-fit residual matrix, however, the previous implementation could not coherently sample whole-unit residual vectors—some columns had NA cells that broke the empirical resampling. The current implementation switches to a unit-level Gaussian approximation in this case: residuals are drawn from a Gaussian fitted to each unit's residual structure rather than from the empirical pool, naturally accommodating missing cells. The fix is scoped to GSC with the parametric bootstrap on data with missing cells; balanced panels see no change in behavior.

Migration. Code from prior `gsynth` releases (March 2017 through v1.3.0) that combined IFE-EM with the parametric SE will return the v2.2.0 hard-gate error; the recommended replacement is `vartype="bootstrap"` or `vartype="jackknife"` when N_{tr} permits. Code that relied on block CV's defaults will continue to work but may yield different rank picks under v2.3.0 rolling CV; pass `cv.method="block"` to reproduce the prior behavior.