INNOCENZO GASPARINI INSTITUTE FOR ECONOMIC RESEARCH

Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

Do Not Drain The Swamp! Populism, Bureaucracy and Economic Performance

Massimo Morelli, Dmitrii Petrukhin, Matia Vannoni

Working Paper n. 730

This Version: June 20, 2026

IGIER – Università Bocconi, Via Guglielmo Röntgen 1, 20136 Milano –Italy
http://www.igier.unibocconi.it

Do Not Drain The Swamp! Populism, Bureaucracy and Economic Performance*

Massimo Morelli† Dmitrii Petrukhin‡ Matia Vannoni§

June 20, 2026

Abstract

Populist leaders often campaign against bureaucratic constraints, yet their economic performance may depend on the very institutional environment they inherit. We examine whether bureaucracy moderates the economic costs of populism using two complementary settings: cross-country synthetic-control evidence and a U.S. state-year panel spanning 1929–2023. The cross-country analysis offers suggestive evidence that the output decline following populist takeovers is attenuated in countries with stronger pre-existing bureaucratic institutions, particularly where bureaucracies are more independent. The U.S. state analysis provides sharper, within-country evidence. Leveraging LLM estimates of gubernatorial populism derived from State of the State speeches, we find that populist governors are associated with significantly lower per-capita income in states lacking an independent civil service, while no comparable penalty emerges—and the association is more than offset—where such protections are in force. Heterogeneity-robust event-study estimates point in the same direction, with the post-entry income decline concentrated in states without civil-service reform. Together, the evidence is consistent with independent bureaucracies acting as a partial buffer against the economic costs of populist governance: the "swamp" appears to dampen the damage populism would otherwise cause.

1 Introduction

"Because we are draining the swamp, it's very simple, and the days of rule by unelected bureaucrats are over."
Donald J. Trump, joint session of Congress (Mar 5, 2025)

"Bureaucracy, the judiciary, oversight bodies, agencies, unelected institutions — vital organs of the Republic that, according to some members of the executive branch, sometimes operate like a shadow government to block or slow down ministerial action. 'The interference of the Deep State in our decisions is an obstacle.'"
Giorgia Meloni, *Quel "deep state" contro il governo* — Il Giornale, 21 November 2025

The quotes above are only but a few examples of populist leaders campaigning against the checks and balances that the political system imposes on them: courts, agencies and especially the bureaucracy. The reason for the animosity of populist politicians towards the public administration is both rhetorical and strategic. Firstly, unelected bureaucrats, especially top level civil servants, are seen as part of the technocratic elites or at least they are perceived as being close in spirit to those elites (Weyland, 2001; Müller, 2016). In the populist rhetoric, civil

*We wish to thank seminar participants in Cambridge for useful comments and Bocconi University and the Janeway Institute for financial support. The usual disclaimer applies.

†Bocconi University, IGIER, Baffi, LISER and CEPR
‡University of Cambridge
§King's College London

servants are seen as aloof from the 'virtuous' people (Meier et al., 2019), hence pandering to the ever present elite v. people rhetorical dimension of populism (Mudde, 2004; Mudde and Kaltwasser, 2013). Secondly, this animosity is a strategic device: expert bureaucrats (especially those chosen by other administrations or those that have worked under those administrations) represent an obstacle to the populist politician's agenda, which is unconditional, and based on simple commitments that voters can easily assess (Bellodi et al., 2023). Bureaucrats can slow-down or even stop the populist political agenda, based on their expertise (Sasso and Morelli, 2021; Bellodi et al., 2023) or even based on their own political preferences (Bellodi et al., 2026).

Building on this, we study whether the economic consequences of populist governments vary with checks and balances, with a particular focus on the independence and capacity of the bureaucracy. We do so with two empirical strategies. First, we replicate the analysis in Funke et al. (2023), based on the balanced core sample of 28 populist episodes (drawn from their broader database of 51 populist national leaders) from 1900 to 2020. We add to this dataset information on the independence and capacity of bureaucracy from V-Dem (V-Dem Institute, 2026). Using the same SCM approach, we find that an independent and strong bureaucracy mediates the negative economic effects of populist takeovers. The mediating effects seem to be driven by bureaucratic independence, rather than capacity.

Second, we test the effect of bureaucracy on the link between populism and economic performance in the context of the U.S. states, because of their similarity in terms of political systems, bureaucracies and economies. The U.S. state analysis relies on a state–year panel (1929–2023) constructed from State of the State speeches and administrative data.

We assemble a corpus of 3,355 gubernatorial speeches and measure populism with LLMs and the holistic-grading prompt of Tamaki et al. (2025). We complement this with data on party affiliation, civil-service reforms, and per-capita income. We find that populist governors are associated with lower per-capita income in states where an independent civil service is not present, in line with the cross-country evidence by Funke et al. (2023). This negative association is attenuated, and in the main specification more than offset, where civil service reforms are in place. The analysis is robust to: different measures of populism, including dictionary-based scores following Gennaro et al. (2024) and Pauwels (2011), to using a different LLM, and to alternative threshold definitions of populism; different measures of civil service reform; additional controls including lagged dependent variable, different functional forms of the outcome variable and coefficient stability Oster (2019) bounds.

For causal evidence, we estimate a heterogeneity-robust dynamic DiD of de Chaisemartin and D'Haultfoeuille (2024) in the post-2010 sample, where civil-service status is stable within the event-study window; the results point in the same direction, with per capita income decline after populism takeover concentrated in states without civil-service reform in force.

Our findings bring together the strands of the political science and political economy literature that study the effects of populism on economic performance and those that look at how populism interacts with the bureaucracy, and checks and balances in general. There is evidence both at the local (Bellodi et al., 2024) and national level (Funke et al., 2023) that populism has negative effects on economic performance. Moreover, the literature documents a tension between populism and bureaucracy, and checks and balances in general. Populist politicians campaign against bureaucrats, courts and unelected entities (Peters and Pierre, 2019; Bauer and Becker, 2020), and once in power, they try to dismantle those checks and balances (Sasso and Morelli, 2021; Bellodi et al., 2024; Kyriacou and Trivín, 2025). In the paper, we find that those very institutions that populist governments try so hard to destroy instead provide crucial checks and balances that attenuated the governments' worst inclinations and their effects on the economy.

2 Literature Review

This paper sits at the intersection of three strands of the literature: the economic consequences of populism, the relationship between populism and public administration, and the role of checks and balances, in particular bureaucratic institutions, in moderating the interaction between electoral and economic outcomes.

Populism and economic performance. A large and growing body of evidence documents the negative economic consequences of populist governance. The most direct predecessor of our cross-country analysis is Funke et al. (2023), who construct a dataset of 51 populist leaders across the world from 1900 to 2020 and apply synthetic control methods to estimate the causal effect of their takeover on real GDP per capita and other macroeconomic outcomes. They find that, fifteen years after a populist takeover, GDP per capita is on average around 10% below the synthetic counterfactual, with additional deterioration in trade openness, foreign investment, and democratic quality. Earlier work traced a recurring macroeconomic cycle characteristic of populist governance in Latin America: expansionary fiscal policy generates short-run growth but ends in high inflation, balance-of-payments crisis, and output collapse (Dornbusch and Edwards, 1990). Similar results are presented in Magud and Spilimbergo (2021). More recent evidence has extended this picture to the local level. Bellodi et al. (2024) exploit close mayoral elections in over 8,000 Italian municipalities across two decades and find that electing a populist mayor leads to worse fiscal performance, a larger share of procurement contracts with cost overruns, and a sharp rise in bureaucratic turnover and decline in bureaucratic expertise. Our U.S. state-level analysis complements this local evidence by providing estimates of the economic costs of populism at the subnational level in the United States, exploiting variation in both leadership style and pre-existing institutions.

Populism, public administration, and institutional legacies. A related literature examines the institutional consequences of populist governance and, in particular, how pre-existing institutional strength shapes the damage that populists can inflict. Peters and Pierre (2019) argue that the anti-elitist logic of populism is fundamentally at odds with a professionalized bureaucracy: populists view expert career officials as members of the same corrupt elite they claim to oppose, creating a structural hostility toward merit-based public employment. Bauer and Becker (2020) document the administrative strategies pursued by populist governments across a range of historical cases, showing that purges of senior career officials and the replacement of merit-based appointments with political loyalists are a recurring feature of populist governance in office. Sasso and Morelli (2021) provide a theoretical account of this dynamic, showing how independent bureaucrats can block or delay implementation of the populist agenda on grounds of expertise or legality, and how populist leaders respond strategically by seeking to undermine civil-service protections. Bellodi et al. (2024) provide causal evidence consistent with this mechanism: populist mayors in Italy are associated with significantly higher forced turnover among top municipal bureaucrats and a sharp reduction in the graduate share of the civil service.

The question of whether pre-existing institutional quality can limit the damage done by populists has begun to receive direct empirical attention. Kyriacou and Trivín (2025) apply synthetic control methods to the same set of populist episodes used by Funke et al. (2023) and ask whether the erosion of the rule of law after a populist takeover depends on institutional legacies at the time of the takeover. They find that, in the aggregate, populism reduces V-Dem's rule of law index by around 11 percentage points after fifteen years. Crucially, however, the effect is concentrated in countries with a weak pre-existing rule of law tradition: in low-legacy countries the decline reaches 17.5 percentage points, while in high-legacy countries it is only 5.8. This is precisely the logic our paper pursues for economic performance and bureaucratic independence: institutional pre-conditions at the time of the populist takeover determine how

much damage is done. We build on Kyriacou and Trivín and expand in three main ways. First, we study the effect of populist takeover on economic outcomes, in line with the strand of literature discussed above. Second, our focus is on bureaucratic independence, measured via civil-service reform in the U.S. and a composite of V-Dem bureaucracy indicators in the cross-country analysis, rather than the rule of law tradition as a whole.[1] In this way, we manage to link the discussion on the effects of populism on the economy with that on the tension between populist governments and bureaucracy. Third, we augment our cross-country analysis with U.S. state-level panel, thus exploiting within-country variation. Because U.S. states share a common political system, legal framework, and macroeconomic environment, comparing states with and without civil-service reform holds constant a wide range of confounders that are more difficult to control for in cross-country settings. Moreover, the civil service reforms in the period under study are arguably exogenous reforms imposed by the federal level. The two empirical strategies are therefore complementary: the cross-country analysis establishes the generality of the pattern across a wide range of historical populist episodes, while the U.S. panel provides sharper within-country identification. Together they also speak to external validity: the fact that both approaches point to the same conclusion strengthens confidence that the moderating role of bureaucratic independence is a robust feature of the data rather than an artefact of any particular empirical design.

Checks and balances, bureaucracy, and economic outcomes. Beyond the populism literature, our paper connects to a broader body of work on how checks and balances, and bureaucratic institutions as one key component, moderate the link between political governance and economic performance. At the macro level, Besley et al. (2022) review the economics of bureaucracy and development, documenting a robust positive relationship between bureaucratic quality and GDP per capita both in the cross-section and within countries over time. Within the United States context, Ash et al. (2022) study the political economy of civil-service reform in U.S. states, showing how the adoption of merit-based hiring rules insulates public employment from political patronage: the key source of variation we exploit in our state-level analysis. Finally, a broader literature on the interaction between leader quality and institutional constraints points in the same direction. Jones and Olken (2005) exploit the deaths of national leaders as a source of exogenous leadership change and find that leaders have particularly large effects on growth in autocratic settings. Clark et al. (2014) reach an analogous conclusion in the corporate context, showing that CEOs matter most when ownership and governance structures are weak or ambiguous. Ottinger and Voigtländer (2025) provide the sharpest evidence, exploiting plausibly exogenous variation in the cognitive ability of European monarchs generated by centuries of elite inbreeding: ruler quality has large effects on state performance, amplified precisely when institutional checks are weak. This body of evidence directly anticipates the logic of our paper: just as leader quality matters more when institutional constraints are weak, a populist leader causes more economic damage when bureaucratic institutions are too weak to constrain the implementation of the populist agenda.

3 Cross-Country Evidence

Funke et al. (2023) provide cross-country evidence that the election of populist leaders is followed by economically meaningful medium-run declines in output and other macro-outcomes. In this part, we test whether this effect is attenuated by the presence of a strong and independent bureaucracy, using the latest V-Dem data (v16) (V-Dem Institute, 2026).

[1] The V-Dem rule of law index used by Kyriacou and Trivín (2025) consolidates judicial constraints on the executive, access to justice, and political corruption. It does not include any measure of bureaucracy, which is what we use for the cross country analysis.

3.1 Data

We study the GDP consequences of populist leadership using the core GDP synthetic-control sample from Funke et al. (2023). This sample contains 28 populist takeovers for which the GDP doppelgänger exercise can be estimated over a balanced event window from $t = -15$ to $t = 15$. Funke et al. (2023) also report results for an extended sample of 51 episodes with shorter, unbalanced event windows; we restrict attention to the balanced core sample so that the same set of episodes contributes to the high- and low-bureaucracy averages at every event-time horizon.[2] The outcome is log real GDP per capita, normalized to zero in the event year, so treatment effects are interpreted as log-point deviations from the synthetic counterfactual.

Our institutional moderator is bureaucratic independence and capacity, measured using V-Dem v16 (V-Dem Institute, 2026). We construct a bureaucracy index as the first principal component of three pre-treatment V-Dem indicators: rigorous and impartial public administration (v2clrspct), merit-based state recruitment (v2stcritrecadm), and bureaucratic remuneration (v2strenadm). The sign of the component is oriented so that larger values indicate stronger bureaucratic independence and capacity. We use a transparent PC1 composite of these already-aggregated country-year inputs rather than V-Dem's official latent-variable bureaucracy aggregation because we do not observe the underlying coder-level scores required to replicate V-Dem's full Bayesian measurement procedure; details, including pairwise correlations and validation, are in Appendix F.2.

We classify populist episodes according to bureaucratic variable in the year before the populist leader takes office. The cases above the median among usable treated events are classified as high-bureaucracy cases, and cases at or below the median are classified as low-bureaucracy cases.[3] The GDP SCM is estimated for all 28 treated episodes. The smaller sample in the bureaucracy-split results is due to the split requiring a valid pre-treatment bureaucracy score measured in the year before treatment ($t-1$). The composite bureaucracy index is observed only when all three V-Dem components are available (rigorous and impartial administration, merit-based state recruitment, and bureaucratic remuneration). Four core cases lack a valid value at $t-1$: Slovakia 1990, Peru 1985, Peru 1990, and Japan 2001—and are therefore excluded from the split, yielding $N = 24$ for the split-sample analysis.

3.2 Empirical strategy

Our cross-country analysis uses the synthetic control method (SCM) proposed by Abadie and Gardeazabal (2003), Abadie et al. (2010), and Abadie (2021), following the implementation in Funke et al. (2023). The method constructs, for each populist episode, a data-driven counterfactual: a synthetic doppelgänger, from a weighted combination of donor countries that did not experience populist leadership during the relevant window. The weights are chosen to minimize the pretreatment distance between the treated country and its synthetic counterpart, matching on the trajectory of real GDP per capita. Comparing the treated country's post-treatment path to that of its doppelgänger yields the synthetic-control gap, which measures the effect of the populist treatment at each post-treatment horizon.

More formally, for each treated episode $p = 1, \ldots, P$ occurring in country c at year T_p, let

[2] Two features of the extended sample motivate this choice. First, extended episodes with starting years 1919–1938: Hitler 1933, Mussolini 1922, Vargas 1930, Cárdenas 1934, and Velasco 1934 have post-treatment trajectories that run into World War II. Second, extended episodes with starting years 2005–2019: Bolsonaro 2019, AMLO 2018, Trump 2017, Modi 2014, among others are right-censored within our estimation window, so the set of episodes contributing to the high–low average would vary across horizons τ. Restricting to the core sample keeps the contributing set fixed at all $\tau \in [-15, +15]$ and makes the permutation-based randomization test well-defined.

[3] We split events at the median rather than using an alternative threshold because it yields balanced high- and low-bureaucracy groups. The sample available for the heterogeneity analysis is smaller than in Funke et al. (2023) and Kyriacou and Trivín (2025) due to missing values on the V-Dem bureaucracy variables for some episodes.

$\mathbf{Y}_p$ denote the vector of pretreatment covariates in the treated country and let $\mathbf{X}_p$ denote the corresponding matrix of covariates for the C countries in the donor pool (excluding countries that themselves experienced populist leadership during the relevant window).[4] Let $\mathbf{W}_p = (w_1^p, \ldots, w_C^p)'$ denote the vector of donor weights. The optimal weighting vector $\mathbf{W}_p^*$ is chosen to solve

$$\min_{\mathbf{W}_p} \left(\mathbf{Y}_p - \mathbf{X}_p\mathbf{W}_p\right)' \mathbf{V}_p \left(\mathbf{Y}_p - \mathbf{X}_p\mathbf{W}_p\right), \quad p = 1, \ldots, P, \tag{3.2.1}$$

subject to $\sum_{c=1}^{C} w_c^p = 1$ and $w_c^p \geq 0$ for all c, where $\mathbf{V}_p$ is a positive semidefinite and symmetric matrix whose elements are selected using the data-driven approach of Abadie et al. (2010). The matrix $\mathbf{V}_p$ governs the relative importance of each covariate in the matching objective.

For each treated episode and event time $\tau \in \{-15, \ldots, 15\}$, where $\tau = 0$ denotes the takeover year, let $Y_{p\tau}^{\text{act}}$ denote the observed outcome and $Y_{p\tau}^{\text{sc}}$ the synthetic-control outcome constructed using $\mathbf{W}_p^*$. The episode-level gap is

$$G_{p\tau} = Y_{p\tau}^{\text{act}} - Y_{p\tau}^{\text{sc}}.$$

where $Y_{p\tau}^{\text{act}}$ is the observed outcome and $Y_{p\tau}^{\text{sc}}$ is the corresponding synthetic-control outcome. Averaging these gaps across treated episodes yields the aggregate effect of populist governance on the outcome at each horizon, as in Funke et al. (2023). For further detail, please refer to Funke et al. (2023).

To study heterogeneity, we split treated episodes into high- and low-institution groups using the institutional value in the year before treatment. Specifically, for each institutional measure B and treated episode p, we compute the lagged value B_{p,T_p-1} and compare it to the median lagged value among treated episodes in the estimation sample. An episode is classified as *high* if

$$B_{p,T_p-1} > \text{Median}\left(B_{q,T_q-1}\right)_{q\in\mathcal{I}},$$

and as *low* otherwise, where $\mathcal{I}$ indexes treated episodes with usable estimates. All splitters are coded so that higher values correspond to stronger institutions or checks and balances.

Our split variables focus on a composite bureaucracy measure constructed from three V-Dem indicators, rigorous and impartial administration, merit-based state recruitment, and bureaucratic remuneration, as well as each component individually. The construction of the composite is described in Appendix F.2. In other words, these variables measure bureaucratic capacity and independence.

Bureaucracy split sample. As mentioned above, four core cases do not have a valid bureaucracy value at $t-1$: Slovakia 1990, Peru 1985, Peru 1990, and Japan 2001. Slovakia 1990 lacks the required lagged institutional data, while the Peru and Japan cases are missing one or more bureaucracy components. Because these cases cannot be assigned to either the high- or low-bureaucracy group, they are excluded from the bureaucracy split analysis. The resulting bureaucracy split sample is therefore $N = 24$, while the underlying GDP SCM remains estimated on all 28 Funke core cases.

Within each group, we aggregate episode-level synthetic-control gaps using simple arithmetic means:

$$\bar{G}_\tau^H = \frac{1}{N_H}\sum_{p\in H} G_{p\tau}, \qquad \bar{G}_\tau^L = \frac{1}{N_L}\sum_{p\in L} G_{p\tau},$$

where H and L denote the sets of high- and low-institution episodes, respectively. The high-minus-low treatment-effect difference is then

$$\Delta_\tau = \bar{G}_\tau^H - \bar{G}_\tau^L.$$

[4] Here, as in Funke et al. (2023), we use the pre-treatment outcome variable.

Each successfully estimated treated episode receives equal weight within its group. We prefer this unweighted specification because inverse-RMSPE weighting is overly sensitive to a small number of extremely low-RMSPE cases.[5][6]

To assess whether the observed high–low difference path Δ_τ is larger than would arise by chance, we use a permutation-based randomization procedure inspired by the placebo-based ATT inference framework in Abadie et al. (2010). Under the sharp null that the institutional classification is uninformative, the episode-level gap paths $G_{p\tau}$ are exchangeable across the "high" and "low" labels. We therefore repeatedly reshuffle the high/low labels across treated episodes (preserving the original group sizes), recompute Δ_τ for each reshuffle, and obtain a null distribution at each event time τ. In the figures, the solid line plots the observed Δ_τ and the shaded region shows the pointwise 5th–95th percentile envelope of the permutation distribution (a 10% randomization band). Details are provided in Appendix F.5.

3.3 Results

Figure 1 summarizes our main macro pattern. Panel (a) plots average GDP-per-capita trajectories for treated episodes and their synthetic controls, splitting episodes by whether the composite bureaucracy measure lies above or below its pre-treatment median. Panel (b) plots the high-minus-low difference in average synthetic-control gaps, together with the pointwise 5th–95th percentile permutation envelope under the null of no heterogeneity (a 10% randomization band).

Before turning to the main findings, we note several features of the evidence that support a causal interpretation. First, both groups track their synthetic counterfactuals closely in the pre-treatment period, confirming the quality of the synthetic matches. Second, and more importantly, the high-minus-low difference in average gaps is close to zero throughout the pre-treatment window: the two groups were on parallel trajectories before the populist takeover, so the post-treatment divergence cannot be attributed to pre-existing differences in economic performance between high- and low-bureaucracy episodes. Third, the results survive a time-placebo test in which the treatment date is shifted five years earlier: the high-minus-low difference remains inside the permutation envelope during the placebo window, suggesting that the post-takeover divergence is not a statistical artefact of the estimation procedure. Finally, our permutation-based inference directly addresses the concern that the high–low split might capture any salient grouping of episodes rather than bureaucratic capacity specifically: across 1,000 random reshufflings of the high/low labels (preserving the original 12/12 group sizes), the observed post-takeover difference exceeds the vast majority of differences generated under the null, indicating that the bureaucracy classification carries genuine information.

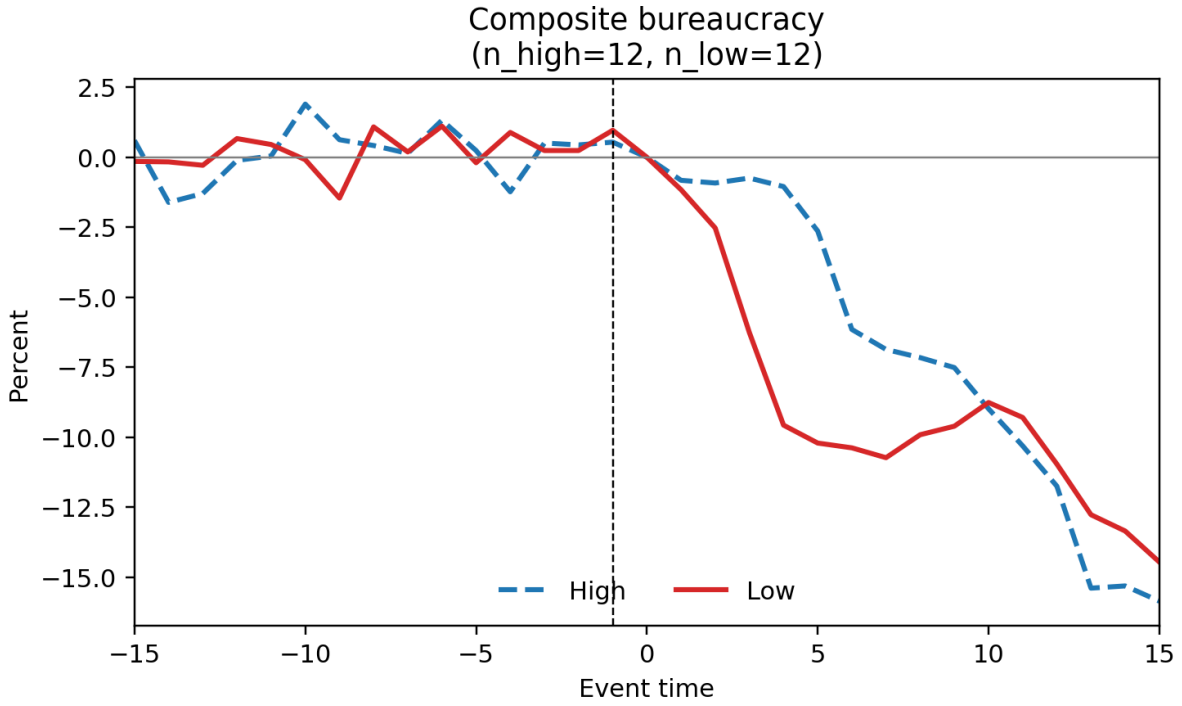
The results suggest short-run attenuation but little evidence of a persistent difference over the full medium-run horizon. The high-minus-low difference peaks at $\tau = 4$ (four years after populists enter office): GDP is 0.010 log points below its synthetic counterfactual in high-bureaucracy cases, compared with 0.096 in low-bureaucracy cases — a gap of 0.085 log points (the displayed figures are rounded; the difference is computed from the unrounded gaps). Using permutation inference that randomly reshuffles the high/low bureaucracy labels across treated episodes while preserving the 12/12 group sizes, the pointwise randomization p-value at this horizon is 0.054, and the observed difference lies just outside the 90% reshuffled-null envelope of

[5] Across the 28 core GDP events, Argentina 1946 alone receives 80.2 percent of the total inverse-RMSPE weight, while Argentina 1946, Brazil 1951, and Slovakia 1990 together receive 99.99 percent. In the bureaucracy split, the low-bureaucracy average is effectively determined by two cases: Argentina 1946 receives 87.9 percent of the group weight and Brazil 1951 receives 12.1 percent. The high-bureaucracy average is also highly concentrated, with Italy 2001 receiving 67.5 percent and Italy 1994 receiving 22.7 percent of the group weight. For this reason, inverse-RMSPE aggregation would no longer represent the average experience of high- and low-bureaucracy populist episodes. It would instead compare a small number of exceptionally well-fitting cases.

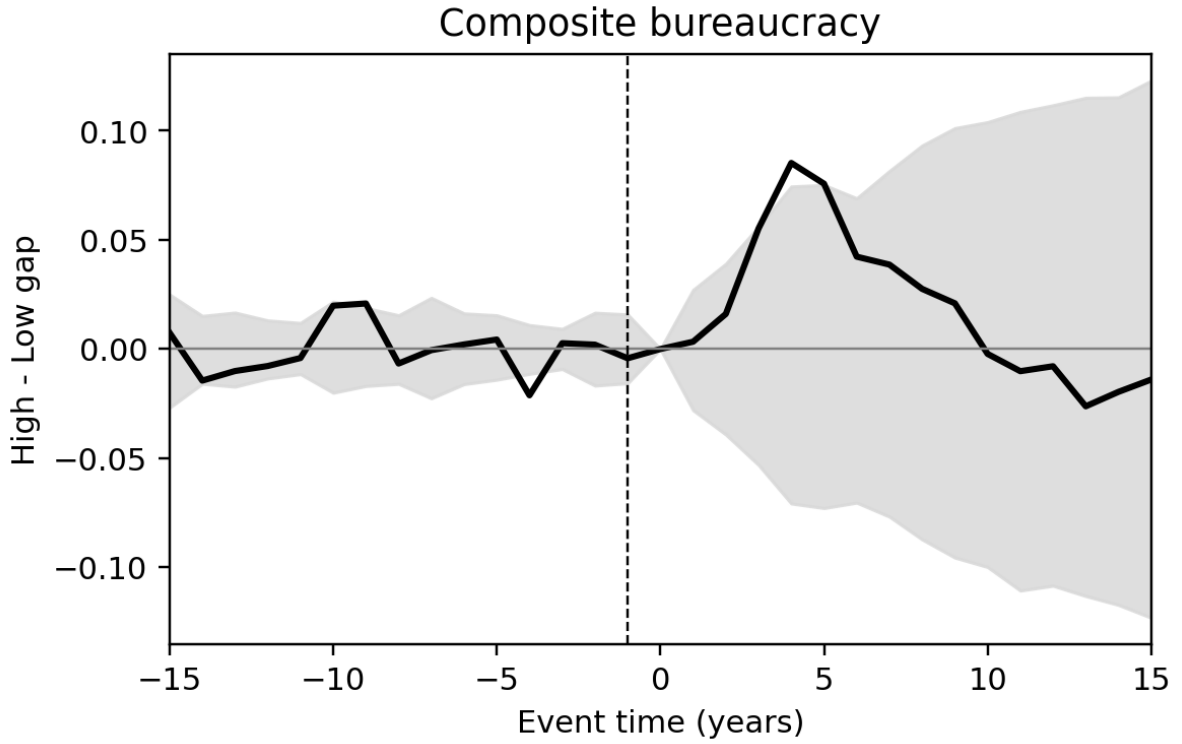
[6] Funke et al. (2023) use simple averages; Kyriacou and Trivín (2025) use inverse-RMSPE weights. We follow the former.

[−0.071, 0.074]. By $\tau = 10$ and $\tau = 15$, however, both groups exhibit sizeable GDP losses and the high–low difference narrows toward zero. Stronger bureaucratic capacity therefore appears to delay or soften the initial output cost of populism, but the attenuating effect is not sustained over the full medium-run horizon we consider.

This is consistent with the findings in Kyriacou and Trivín (2025), who show that populist governments, especially where they inherit strong checks and balances, start eroding those democratic checks and balances after five years of getting into power. Indeed, if populists start eroding checks and balances, including a strong and independent bureaucracy, we will see the moderating effects of the bureaucracy disappear over time.



(a) GDP-per-capita trajectories for treated episodes and their synthetic controls, split by pre-treatment composite bureaucracy (high vs. low).



(b) High-minus-low difference in average synthetic-control gaps. The shaded region shows the pointwise 5th–95th percentile permutation envelope under the null that institutional labels are uninformative; where the observed path exits the envelope, the null of no heterogeneity is rejected at the 10% level.

Figure 1: Heterogeneity in post-takeover GDP-per-capita gaps by pre-treatment composite bureaucracy. Panel (a) plots average trajectories for high- and low-bureaucracy episodes against their respective synthetic controls. Panel (b) plots the high-minus-low difference in average gaps with permutation-based inference. The two panels show the same comparison from complementary angles.

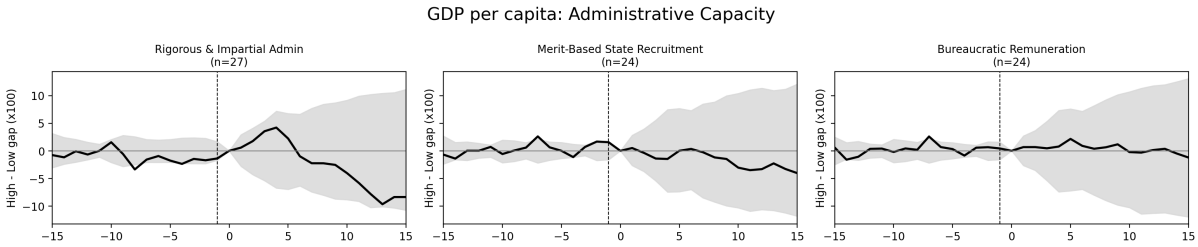


Figure 2: Bureaucratic capacity. The figure reports high–low differences in average synthetic-control GDP-per-capita gaps for the bureaucracy/state-administration measures and their components. The shaded region shows the pointwise 5th–95th percentile permutation envelope under the null of no heterogeneity.

The composite result suggests that bureaucratic independence and capacity attenuate the short-run economic costs associated with populist takeovers. To unpack this finding, Figure 2 examines each component of the composite index separately. The pattern appears to be driven primarily by rigorous and impartial administration, although the high-low difference for this component, like for the other two components, does not exit the permutation envelope at conventional levels. We therefore read the component breakdown as suggestive rather than conclusive: of the three components, rigorous and impartial administration shows the largest post-takeover divergence between high- and low-bureaucracy episodes, consistent with bureaucratic *independence* (rather than remuneration or recruitment criteria alone) being the relevant margin, but the per-component sample sizes are too small to formally reject the null of no heterogeneity for any single component. This pattern parallels our U.S. state-level findings, where institutional-rule measures (civil service reform) produce sharp results.

These results come with some limitations: the number of treated episodes is small (12 per group for the composite split, i.e., $N = 24$ because four of the 28 core episodes lack a valid bureaucracy value at $t-1$; and 24–27 per component split depending on component availability), and sample sizes of this magnitude limit statistical power. We therefore treat the macro patterns as descriptive evidence that motivates the micro-level analysis in the next section, where we have within-country variation and larger samples, as well as similar political and economic contexts.

4 Micro Evidence: The Case of the United States

In this section, we seek to corroborate the cross-country findings by looking at the U.S. states. U.S. states represent a valid setting to test our expectations on the link between populism, bureaucracy, and economic outcomes. Broadly speaking, US states present relatively similar political, economic and social features, especially throughout the 20th century. Moreover, focusing on U.S. states allows using the same definition of independent bureaucracy. We look at the extension to U.S. state agencies of the principles of an independent bureaucracy already applied at the federal level. Not only does this make these reforms comparable across U.S. states, but they are arguably exogenous to the politics and economy of the US states, as they were imposed by the federal level, especially the establishment of the civil service, less so its repeal. Also, by focusing on U.S. states, we can extract populist rhetoric from relatively similar speeches by politicians from similar parties.

4.1 Case Studies: LePage and Blagojevich

Before describing the data and our general analysis, it is useful to introduce two illustrative cases: they are examples of populist governors who both committed to policies with negative economic consequences, but where the existing bureaucracy could limit the damage. We illustrate these

three components in turn.

Populist policies. Paul LePage's governorship in Maine from 2011 to 2019 provides a clear case of populist executive politics translated into policy. Across successive State of the State addresses, LePage framed economic and social policy as a struggle between ordinary Mainers and self-interested political, bureaucratic, and sectoral elites, located in the capital Augusta or Washington. In 2013, he argued that government had not strengthened Maine families, but instead had taken "more and more of our family's hard working income away to serve some people's political and/or financial self interests" (LePage, 2013). The same speech presented transparency and anti-waste initiatives as a direct attack on misused public money: "We exposed the wasteful use of Mainers tax dollars at agencies such as the Maine Turnpike Authority and Maine State Housing Authority. We not only exposed it–we cleaned it up" (LePage, 2013). This anti-elite style became more explicit in later years. In 2016, LePage denounced "socialist politicians in Augusta," "career politicians and their allies in media," and claimed that they had criticized his administration while "the Maine people are tired of the games" (LePage, 2016). He also framed energy policy in distributive and anti-special-interest terms, arguing that "the Maine people deserve a break, not the wealthy special interests in Augusta," and that subsidies for wind and solar benefited "only a few wealthy investors" (LePage, 2016). In 2017, this became a sharper ideological attack: "Liberals are now trying to transform our state into a socialist utopia," while the solar industry would "line its pockets on the backs of hardworking Mainers" and "our poor and most vulnerable" (LePage, 2017).

These claims were not merely rhetorical. They were linked to concrete policies: drug testing for a subset of TANF applicants, an anti-tax and anti-government fiscal agenda, and a moratorium on new wind-energy permitting (Worland, 2015; Miller, 2018). In each case, LePage justified policy as defending ordinary taxpayers or vulnerable Mainers against political insiders and wealthy lobbyists. This rhetoric also extended to the administrative state itself. LePage repeatedly presented bureaucracy and the public-sector in general as obstacles to popular government: in 2014 he celebrated that his administration had "reduced bureaucratic red tape," "right-sized government," and "found efficiencies within state agencies," while in 2018 he contrasted his reforms with "politics as usual" and claimed that state employees should not have to subsidize a union "political agenda" (LePage, 2014, 2018).

Illinois provides another example. Rod Blagojevich, who served as the 40th governor of Illinois from 2003 to 2009, presented himself as a direct advocate for ordinary families against Springfield insiders, while attacking established institutions and using executive power aggressively. In the 2008 State of the State speech, the governor emphasised that "by cutting spending and eliminating pork, we can afford to give the people more of their money back" and that "we have to consolidate administrative functions and close unnecessary and wasteful facilities" (Blagojevich, 2008). His administration became associated with politicized appointments and pressure on state boards and agencies that eventually led to Blagojevich's indictment. Federal prosecutors later described a "wide-ranging scheme to deprive the people of Illinois of honest government," alleging that Blagojevich used state appointments, business decisions, legislation, and pension-fund investments to seek campaign contributions, employment, or other financial benefits for himself and his allies (U.S. DOJ, 2009). The Illinois case therefore captures the attempted personalization and politicization of the administrative state itself.

The economic consequences of these policies. The potential costs of LePage's populist governing style are clearest in the renewable-energy domain. Other initiatives followed a similar logic: the TANF drug-testing rule risked increasing administrative burdens and benefit loss among poor households, while the anti-tax agenda raised the possibility that optimistic political claims about growth and efficiency could displace more cautious revenue planning (Moynihan et al., 2015; Worland, 2015). The wind-energy moratorium provides the sharpest illustration of

how populist policy could threaten economic performance. LePage repeatedly framed renewable energy not as a technocratic question of energy mix, prices, and investment, but as a distributive conflict between ordinary ratepayers and politically connected green-energy interests. In 2016, he argued that "socialists love to subsidize new wind and solar energy projects because they think it will save the earth," but that this "expensive and inefficient energy benefits only a few wealthy investors" (LePage, 2016). In 2017, he sharpened the claim further, arguing that "ratepayers are being charged twice" so that households with solar panels could "recoup their money faster," and that "the wealthy solar industry will line its pockets on the backs of hardworking Mainers" and "our poor and most vulnerable" (LePage, 2017).

This rhetoric culminated in the 2018 moratorium on new wind-energy permits, which introduced uncertainty into a sector dependent on predictable permitting rules, long time horizons, and stable expectations about state energy policy (Miller, 2018). The moratorium applied to large parts of western and coastal Maine, coastal islands, and significant migratory pathways, and it prohibited state agencies from issuing permits related to wind turbines while a new advisory commission reviewed wind power's economic effects (Miller, 2018). The timing made the economic stakes especially visible: the order was issued just before Massachusetts regulators were expected to choose renewable-energy suppliers, with several Maine-based wind projects competing for those contracts (Miller, 2018). Industry representatives therefore interpreted the moratorium as a signal to investors and neighboring states that Maine had become a politically risky permitting environment. Jeremy Payne, executive director of the Maine Renewable Energy Association, described it as an attempt to block "billions of dollars of investment" and warned that unilateral executive action could "wreck a billion-dollar industry" (Miller, 2018).

The potential harm was therefore not only environmental. By the time of the moratorium, Maine already had 378 wind turbines and 901 megawatts of installed capacity, more than all other New England states combined, and wind development had created or supported thousands of jobs in the state over the previous decade (Miller, 2018). The moratorium threatened this sector through several channels: delayed capital investment, foregone construction employment, weaker operations and maintenance activity, reduced landowner payments, and lower local tax revenues in host communities. The risk was not hypothetical. The Press Herald article notes that Statoil had already abandoned a $120 million offshore wind demonstration project in Maine in 2013 after LePage's intervention, before later moving forward with an estimated $228 million floating offshore wind project in Scotland (Miller, 2018). The moratorium also created uncertainty for Maine Aqua Ventus, the University of Maine-led floating offshore wind project near Monhegan Island, which was intended to be the first project of its kind in the United States (Miller, 2018). These examples clarify what was at risk: the moratorium did not simply delay turbines; it threatened to divert investment, expertise, and first-mover advantages away from Maine, weakening a clean-energy supply chain that depends heavily on policy credibility and administrative predictability.

The Illinois case illustrates another example of populist policies damaging the government. Blagojevich shows how a populist executive can threaten the quality of economic governance by politicizing personnel, procurement, and public investment. Since the very beginning, he waged a war against a 'corrupt' civil service: "After two months in office, I can tell you that the clean up of State Government is well under way [...] make sure that every state job has only one purpose" (Blagojevich, 2003). The federal indictment alleged that Blagojevich and his associates used state appointments, contracts, grants, regulatory decisions, and pension-fund business as instruments for extracting campaign contributions and private benefits (U.S. DOJ, 2009).

Bureaucratic mediation. The renewable-energy episode in Maine also illustrates how relatively autonomous administrative institutions can mediate the effects of gubernatorial intervention. In January 2018, Governor Paul LePage issued an executive order directing state

agencies not to issue permits for wind turbines in designated areas until a newly created advisory commission had completed its review. Because the order specified no deadline for the commission's report, it generated considerable uncertainty for prospective wind developments, although its legal force and practical implementation were subsequently contested (Miller, 2018; Bever, 2018).

Yet the effects of the intervention were limited by the institutional structure of Maine's energy administration. Efficiency Maine Trust was not an executive department directly controlled by the governor, but a body governed by an independent, multi-member Board of Trustees with fixed terms and subject primarily to oversight by the Maine Public Utilities Commission (Maine Legislature, 2012; Efficiency Maine Trust, 2018). This organisational separation gave the Trust scope to continue implementing its statutory mandate even as LePage sought to halt permits for new wind turbines. Moreover, the executive order targeted wind-energy permitting rather than the Trust's demand-side programmes. Efficiency Maine could therefore continue to administer rebates, technical assistance, energy assessments, consumer information, and support for efficiency investments by households and businesses. During FY2018, which included the introduction of the moratorium, it reported $46.5 million in programme expenditures, more than $47.3 million in associated private investment, 7,780 supported home-efficiency projects, and projected lifetime avoided energy costs exceeding $189 million (Efficiency Maine Trust, 2018). The Trust did not circumvent the moratorium by approving wind projects, nor could its programmes replace the generating capacity and investment associated with utility-scale wind development. Its institutional autonomy nevertheless prevented the governor's intervention in one part of the renewable-energy sector from stopping state-supported energy activity altogether.

This point matters because LePage's own moratorium worked through the permitting agencies: the executive order prohibited state agencies from issuing permits "related to wind turbines," while placing the policy review in a commission whose meetings and documents were exempt from ordinary public-access rules, likely staffed with the governor's 'loyals' (Miller, 2018). The fact that Efficiency Maine was organized as a separate quasi-independent implementation body therefore mattered institutionally: the governor could use executive authority to freeze one permitting channel, but he could not as easily absorb the entire energy-efficiency apparatus into the same politicized review process.

Illinois is useful because the attempted politicization of administration operated in a state with a long-standing merit-system and anti-patronage infrastructure. The Illinois Personnel Code provides the statutory basis for the state's civil-service merit system, covering state employment unless specifically excluded and empowering Central Management Services to administer personnel rules and the Civil Service Commission to monitor the system and conduct hearings (State of Illinois, 2026). The Commission hears appeals of discharges, suspensions, transfers, layoffs, and demotions, and it also reviews exemptions from the classified service for positions involving policy responsibility (Illinois Civil Service Commission, 2026). Illinois was also shaped by the Shakman decrees and related constitutional doctrine, which limited the use of political loyalty as a condition for public employment and made patronage hiring legally contestable. These rules did not prevent Blagojevich-era abuses, especially in exempt appointments, boards, and politically controlled channels. They did, however, limit how far the governor could penetrate the career bureaucracy.

The clearest example is the DeFraties–Casey episode. In 2006, the Blagojevich administration fired two mid-level personnel managers at the Department of Central Management Services after allegations that politically connected job applicants had received improper advantages (STLPR, 2007). The case then moved through the civil-service appeal process. An administrative law judge found that the evidence did not support dismissal and recommended reinstatement after short suspensions; a Sangamon County judge later rejected the administration's case, and the employees returned to their jobs with more than $200,000 in back pay

(STLPR, 2007).Civil-service rules, appeal rights, and independent review made politicized personnel decisions procedurally costly, legally visible, and partly reversible. Blagojevich could place loyalists in exempt or policy-level positions and pressure boards, but he could not as easily replace the whole career bureaucracy with political supporters.

4.2 Data

SOTS speeches and populism scores. Our analysis uses an unbalanced state-year panel built from gubernatorial State of the State (SOTS) speeches. To measure the populism variable, we draw on State of the State (SOTS) speeches collected from `governorspeech.com`.[7] To fill remaining gaps and to cross-check a small number of texts, we additionally use the SOTS corpus assembled in the `sots` repository by `leops95`.[8] We use the raw speech texts from these sources as inputs to the scoring pipeline described below.

All 3,355 speeches are scored for populist content using `qwen/qwen3-235b-a22b-thinking-2507` and the holistic-grading prompt of Tamaki et al. (2025). Each model call supplies a theoretical definition of populism, a detailed rubric with six populist and six non-populist categories, ten anchored example speeches, and a genre instruction alerting the model to the formal, policy-heavy character of gubernatorial addresses. Temperature is set to zero throughout. Of the 3,355 speeches, 3,343 were scored automatically; eight returned valid scores in malformed JSON and were hand-patched; two exceeded the output-token budget; and two failed at the API level but scored successfully on re-run. Full details are provided in Appendix G.

Speech-level scores are aggregated to the governor level by taking the mean across all speeches attributed to a given governor (`llm_pop_mean`). Within-governor variation is modest, so the mean reliably summarizes each governor's rhetorical stance.[9] Governor-level scores are then mapped onto a state-year panel spanning 1866–2023 using the official governor roster of Kaplan (2021), supplemented with manual corrections, post-2020 updates, and fuzzy matching. In transition years, the state-year is assigned to the incoming governor, since SOTS addresses are typically delivered in January or February. Our main binary populism indicator *Pop* equals 1 when `llm_pop_mean` exceeds the 90th percentile of the state-year distribution of the full sample of speeches (cutoff: 0.06) and 0 otherwise.

We validate the Qwen populism scores through two exercises, with full details reported in Appendix G.1. First, we compare Qwen scores against a hand-coded sample of 100 speeches, stratified across score bins to ensure adequate coverage of non-zero cases. The Pearson and Spearman correlations with human scores are 0.420 and 0.404, respectively; 74% of scores lie within 0.1 points of the human rating and 90% within 0.2 points. Second, we re-score the full corpus with an alternative large language model (GPT-5.5) and benchmark both models against the same hand-coded sample. Qwen tracks the human ratings more closely than GPT-5.5 (lower mean error and bias and higher exact agreement), so we use Qwen for the main analysis. We nevertheless report results using the GPT-5.5-based measure as an alternative-measure robustness check; the construction, validation, and results for the GPT-5.5 scores are in Appendix D.2.

Bureaucratic institutions. We use two measures of state-level bureaucratic institutions. Our main measure of bureaucratic institutions is repeal-adjusted civil-service reform, based on the civil-service adoption data in Vannoni et al. (2021) and Ash et al. (2022). This measure

[7] https://governorspeech.com/

[8] https://github.com/leops95/sots

[9] Among the 702 governors observed with more than one SOTS speech, we compute each governor's maximum within person score range, defined as the difference between that governor's highest and lowest Qwen populism score. The median range is 0.0, the 75th percentile is 0.1, and the 90th percentile is 0.2; the maximum is 1.3. Since scores are recorded in 0.1 point increments, most repeated governors exhibit little within person movement, with large changes concentrated in a small upper tail.

equals one when civil-service reform is in force and equals zero before adoption or after an observed repeal. Repeal dates are documented through 2011; for years after 2011, we carry forward each state's 2011 reform status. This coding more closely captures whether merit-system protections were operative in a given state-year, rather than whether a state had ever adopted such protections. As a robustness check, we also report results using the absorbing civil-service reform indicator, coded as one from the year of adoption onward.

Substantively, this variable captures the extension to U.S. state agencies of the principles established by the Pendleton Act, which created a more independent federal bureaucracy Hoogenboom (1959); U.S. Congress, House. Committee on Post Office and Civil Service. Subcommittee on Manpower and Civil Service (1976); Skowronek (1982); Shafritz et al. (2012). These principles include meritocratic recruitment, bureaucratic tenure, and political insulation from patronage. These principles are crucial to test our expectations: civil service reform selects less partisan-driven bureaucrats and imposes weaker political incentives once they are in office. Because our theoretical mechanism concerns whether civil-service protections are actually in force, the repeal-adjusted measure is preferable to a purely absorbing adoption indicator.

Moreover, these reforms, at least those before the 1960s, are arguably exogenous to the internal dynamics of US states. During the nineteenth and early twentieth centuries, state-level reforms were significantly shaped by top-down policy diffusion from the federal government. A key example is the Hatch Act of 1939, which marked a major development in civil service legislation by restricting federal employees' participation in political activities. This legislation subsequently influenced the direction of civil service reforms at the state level. In the same year, Congress also amended the Social Security Act, mandating the establishment of merit-based systems in departments involved in administering the Act.

Outcome and controls. The outcome variable is the log of per-capita personal income drawn from FRED. Governor party affiliation is coded as Democratic, Republican, or Other. Baseline controls include log speech token count and the number of speeches delivered by each governor, included to absorb mechanical differences in speech length and frequency that could affect LLM scores. For the robustness exercises in Section 6.2, we additionally use the Black population share, the foreign-born population share, the number of firms, and the number of bankruptcies, drawn from the Correlates of State Policy Project (Grossmann et al., 2021). These series are sparsely observed in raw annual form; we construct filled state-year series by linear interpolation and extrapolation within state before taking logs. Construction details are in Appendix H.5. We do not include these variables in the baseline specification because of interpolation, however the results are robust to the inclusion of these controls.

4.3 Descriptive Statistics

Sample and missing values. The main analysis panel covers 1929–2023, the period for which the income outcome is available, giving a target grid of 4,750 state-years (50 states $\times$ 95 years). The binary LLM populism indicator *Pop* is observed for 3,861 of these (81.3%), with the 889 unavailable state-years split between 245 panel rows with missing *Pop* and 644 state-years absent from the panel entirely due to lack of usable SOTS speech coverage.[10] Missingness is concentrated in the earlier decades and in states with shorter speech records, but coverage improves substantially over time: from 2000 onwards coverage is near-complete (1,195 of 1,200 possible state-years). The baseline regression sample therefore contains 3,861 state-years across all 50 states from 1929–2023; Appendix H.4 documents attrition and illustrates the missingness structure by year and state.

[10]The underlying SOTS speech/populism data extend further back to 1866; we restrict to 1929–2023 because this is the period over which the income outcome is available.

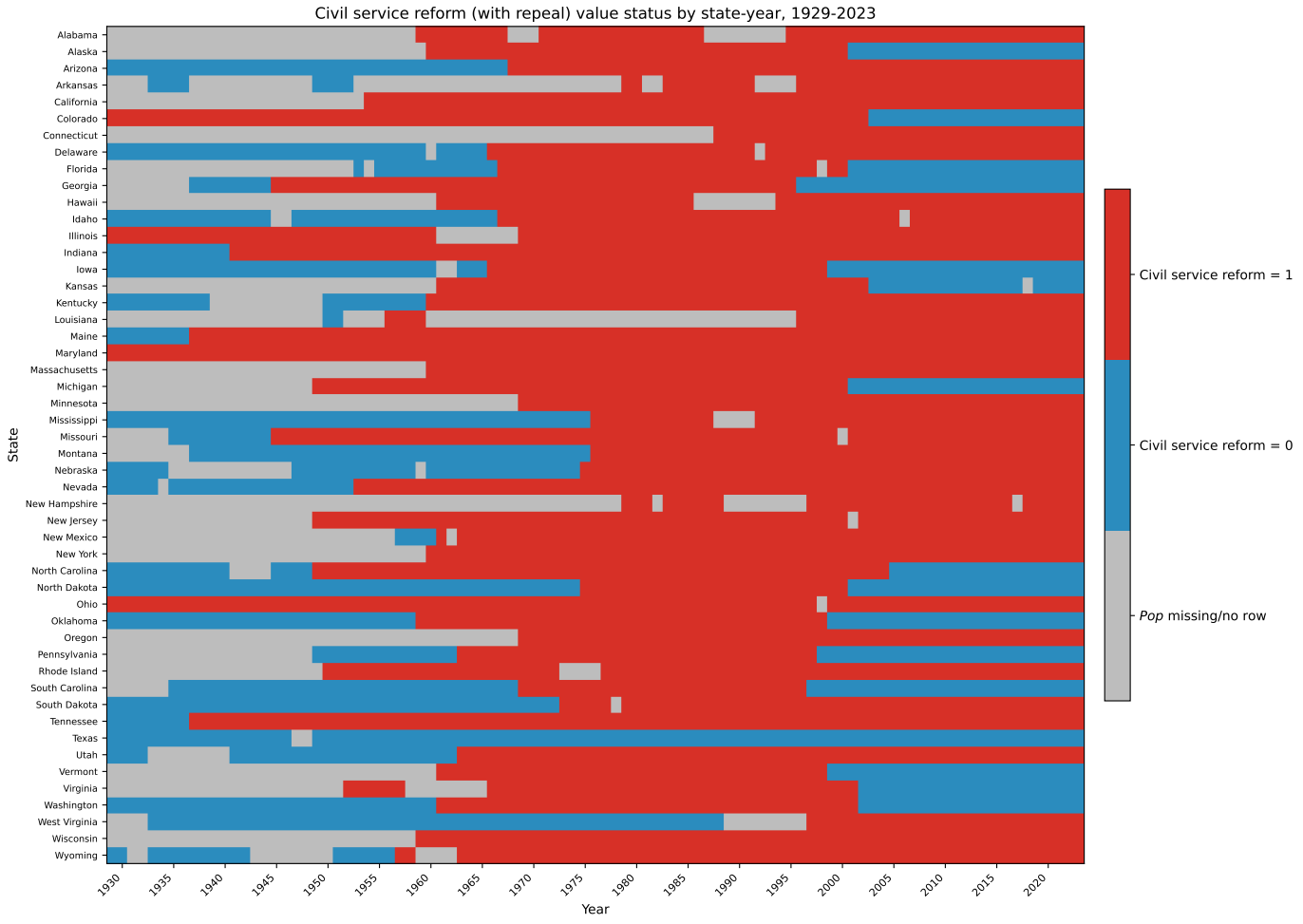


Figure 3: Repeal-adjusted civil-service reform status, by state and year, 1929–2023. Red cells indicate civil-service reform in force after accounting for observed repeals; blue cells indicate either no reform yet adopted or a reform that has been repealed; gray cells indicate state-years with missing populism data.

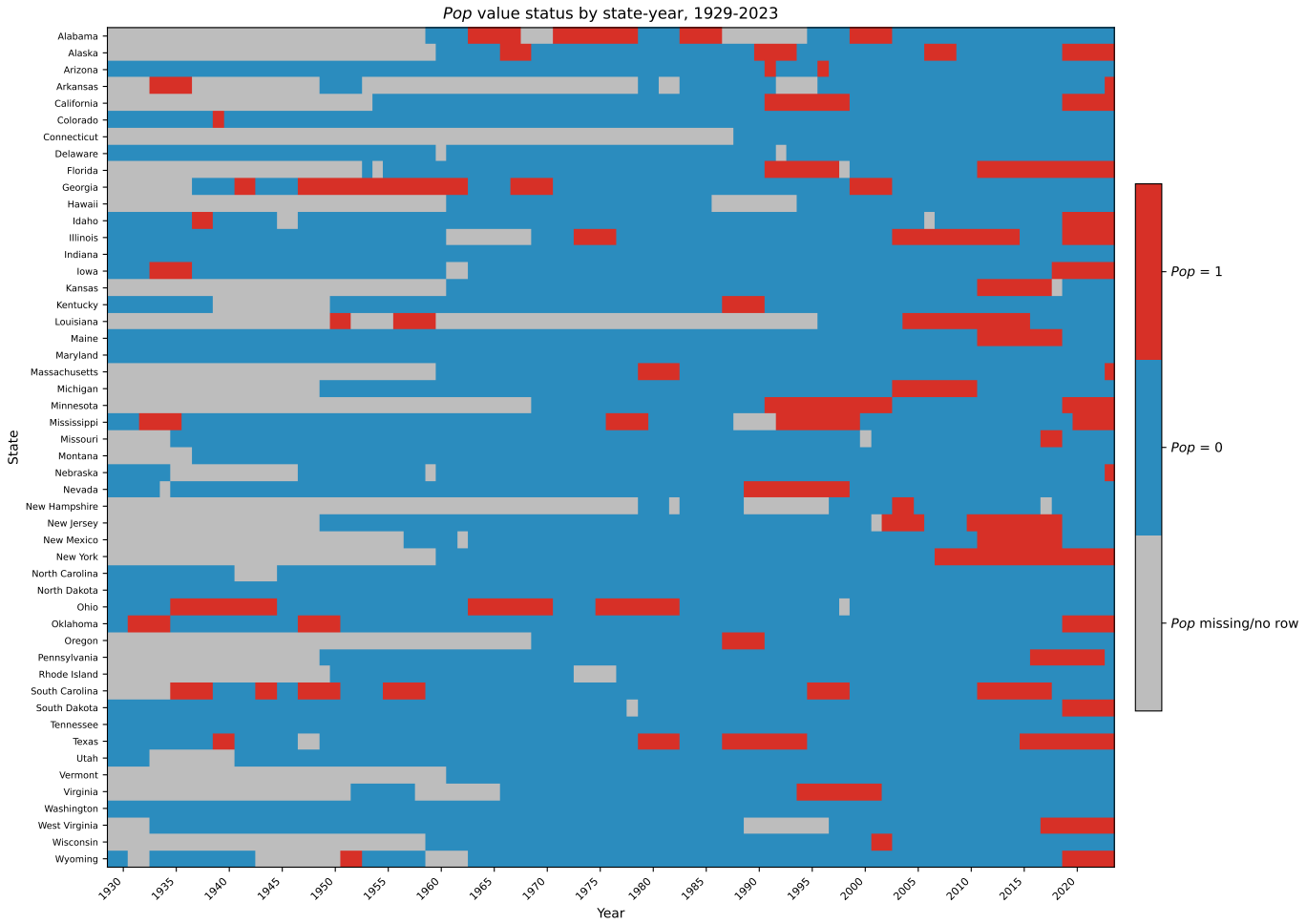


Figure 4: High-populism status, *Pop* by state and year, 1929–2023. Red cells indicate *Pop* = 1 (governor's mean LLM populism score above the 90th percentile of the 1866–2023 state-year distribution); blue cells indicate *Pop* = 0; gray cells indicate state-years with missing values of the populism classification.

The nature of missingness is addressed directly in the empirical specification. State fixed effects absorb time-invariant differences across states, including chronic differences in speech-corpus coverage and long-run political culture. Year fixed effects absorb secular improvements in coverage over time and common shocks. State-specific linear time trends account for the possibility that states with improving coverage are also on distinctive economic trajectories. Identification therefore comes from within-state deviations from each state's own trend, net of common year shocks, rather than from cross-state or cross-period comparisons that would be confounded by the missingness pattern.

As a further check, we show that the results are similar when restricting the sample to 2000–2010, where *Pop* is missing for only three state-years (see Table 9).[11] Lastly, in our DiD approach sample, there are only 2 state-years missing, and the results do not change when accounting for these in different ways as suggested by de Chaisemartin et al. (2025).

Populism. The LLM populism score has a highly skewed distribution at the speech level: 88.3% of the 3,355 scored speeches receive a score of zero, with only 392 speeches (11.7%) receiving a positive score. In the 1929–2023 state-year analysis sample, the governor-level mean

[11] Furthermore, we do not make any assumptions on the bureaucracy measures for this time period.

score has a mean of 0.023 and a maximum of 1.3. Our main binary indicator *Pop*equals one when this mean is strictly greater than the 90th percentile of the full 1866–2023 state-year distribution (cutoff: 0.06), and zero otherwise. Applying this cutoff within the 1929–2023 sample classifies 10.6% of observed state-years as populist. As robustness checks, we also consider a 75th-percentile indicator (24.2% of observed state-years) and a sign-based indicator Pop_{positive} equal to one whenever the mean score is positive (36.3%). The top-scoring speeches in the corpus are consistent with what one would expect from a holistic populism measure. The top 5 highest scored speeches, along with some quotes are provided in Table 1.

Table 1: Highest-Scoring Speeches in the SOTS Corpus

Rank	Governor	State	Year	Score	Representative excerpt
1	Paul LePage	Maine	2016	1.4	"Socialists, career politicians and their allies in media have criticized my Administration every single day [...] Politicians are supposed to represent the Maine people, not special interests, not lobbyists and not a foreign socialist ideology."
2	Lester Maddox	Georgia	1970	1.3	"They're sick and tired of seeing their liberty and their freedom surrendered by public officials [...] They're fed up with public officials who don't care."
3	George Wallace	Alabama	1965	1.3	"A Federal Judge, presiding over a mock court, places a stamp of approval [...] We see today a foreign philosophy that says to the people: you need not bother to work and meet qualifications of a free man — all you must do is demonstrate and cause chaos [...] preservation or destruction of the United States of America."
4	Marvin Griffin	Georgia	1956	1.3	"A tyrannical court ruthlessly seeking to usurp control of state-created, state-developed and state-financed schools [...] Are we going to permit the naked and arrogant declaration of 9 men to destroy our Constitution and to usurp the blood-won rights of our people?"
5	W. Lee O'Daniel	Texas	1941	1.3	"I have seen the hopes of common men wrecked by the crafty maneuvers of these interests [...] a Government of the people, by the people and for the people — all of the people, not just a favored few."

Notes: LLM populism scores are produced by `qwen/qwen3-235b-a22b-thinking-2507` using the holistic-grading prompt of Tamaki et al. (2025). Scores range from 0.0 to 2.0; Excerpts are selected to illustrate populist rhetoric captured by the scoring rubric.

Bureaucratic institutions. Repeal-adjusted civil-service reform is fully coded for all observed state-years in the 1929–2023 panel. Table 6 shows the joint distribution of *Pop* and `cs_ref_with_repeal`. In the baseline regression sample, 408 of 3,861 state-years are classified as high-populism years, and 2,719 state-years have civil-service reform in force.

Other variables. Per-capita personal income ranges from approximately $122 to $89,760 in nominal terms (log range: 4.80 to 11.41). In the baseline regression sample, governors are Democratic in 52.9% of state-years, Republican in 45.8%, and third-party or other in 1.3%. The

average state-year is based on 5.7 State of the State speeches, with an average speech length of 2,837 tokens.

Table 2: Summary Statistics: U.S. State-Year Panel

Variable	N	Mean	Min	Median	Max
Panel A: Outcome					
Log per-capita personal income	4,106	8.983	4.804	9.360	11.405
Panel B: Populism indicators					
LLM score (governor mean)	3,861	0.023	0.000	0.000	1.300
Pop_{75} (score > 75th pctile)	3,861	0.242	0	0	1
Pop (score > 90th pctile)	3,861	0.106	0	0	1
Pop_{positive} (score > 0)	3,861	0.363	0	0	1
Dictionary score (Gennaro et al. (2024))	3,861	0.00316	0.00038	0.00295	0.01638
Dictionary score (Pauwels (2011))	3,861	0.00232	0.00000	0.00215	0.00950
Panel C: Bureaucratic institutions					
Civil service reform	4,106	0.784	0	1	1
Civil service reform (with repeal)	4,106	0.698	0	1	1
Panel D: Controls					
Number of speeches	3,861	5.663	1	5	22
Tokens per speech	3,861	2,837	679	2,480	10,152
Democratic governor	4,106	0.543	0	1	1
Republican governor	4,106	0.445	0	0	1
Other-party governor	4,106	0.012	0	0	1

Notes: The main panel covers 1929–2023, the period over which the income outcome is available. The balanced state-by-year panel would contain 4,750 state-year observations (50 states over 95 years). The observed main panel contains 4,106 state-year rows; *Pop* is observed for 3,861 state-years. *Pop* equals one when the governor's mean LLM score exceeds the 90th percentile of the 1866-2023 state-year distribution (cutoff: 0.06). Dictionary scores are constructed following Gennaro et al. (2024) and Pauwels (2011); see Appendix D.1. Per-capita personal income is in nominal dollars from FRED. Party is coded as Democratic (0), Republican (1), or Other (2); main-panel counts are 2,228, 1,828, and 50, corresponding to shares of 54.3%, 44.5%, and 1.2% in the main panel. The corresponding shares in the baseline regression sample (state-years with the populism indicator observed) are 52.9%, 45.8%, and 1.3%.

5 Empirical strategy

Our empirical strategy tests whether the effect of populist governance on state-level outcomes is moderated by the strength of bureaucratic institutions. The baseline specification is:

$$y_{st} = \beta_1 \,\mathrm{Pop}_{st} + \beta_2 \, B_{st} + \beta_3 \left(\mathrm{Pop}_{st} \times B_{st}\right) + \gamma^\top X_{st} + \delta_s + \lambda_t + \theta_s \cdot t + \varepsilon_{st}. \tag{5.0.1}$$

The unit of observation is a state-year. The outcome y_{st} is the outcome variable–log per-capita income, in state s and year t. The key variable Pop_{st} is a binary indicator equal to one when the governor in office in state s at time t has a mean LLM populism score above the 90th percentile of the full state-year distribution, and zero otherwise.

We operationalize bureaucratic institutions using repeal-adjusted civil-service reform. This measure is based on the civil-service adoption dates in in Vannoni et al. (2021) and Ash et al. (2022). It equals one when civil-service reform is in force in state s and year t, and zero before adoption or after an observed repeal. Since repeal dates are documented through 2011, we carry forward each state's 2011 reform status for later years. This measure captures whether merit-system protections are operative in a given state-year, which is the institutional margin most

closely aligned with our theoretical mechanism: the insulation of bureaucrats from patronage and discretionary political control. As a robustness check, we also report estimates using the absorbing civil-service reform indicator fromVannoni et al. (2021) and Ash et al. (2022), coded as one from the year of adoption onward.

The coefficient of primary interest is β_3, which captures how the association between high-populism governance and income differs when civil-service reform is in force. Funke et al. (2023) document that populist leaders are associated with significant declines in GDP per capita. In the cross-country analysis presented in Section 3, we replicate this finding and show that the negative effect is substantially attenuated in countries with stronger bureaucratic institutions. Our state-level specification tests whether the same logic holds within the United States. A negative β_1 paired with a positive β_3 would indicate that high populism is associated with lower income in states without civil-service reform in force, but that this penalty is attenuated where civil-service protections remain operative. This is the state-level analogue of the macro pattern: bureaucratic independence appears to buffer the economic costs of populist governance.

The vector X_{st} includes time-varying controls: the log token count of the scored speech (to absorb mechanical differences in speech length that could affect LLM scoring), the number of speeches delivered by the governor in office, and a set of party dummies (Democratic, Republican, Other).

We are careful about the inclusion of additional covariates in the baseline specification: because the panel is unbalanced, each additional control variable potentially alters the estimation sample. We therefore keep the baseline parsimonious and defer covariate augmentation to the robustness section. Section 6.2 shows that adding demographic and economic controls does not materially affect the main results.[12]

All specifications include state fixed effects (δ_s), which absorb time-invariant differences across states in long term political culture, institutional design, geography, and economic structure. Year fixed effects (λ_t) control for nationwide shocks: recessions, federal policy changes, political news shocks and secular trends in populist rhetoric, that affect all states simultaneously. We further include state-specific linear time trends ($\theta_s \cdot t$) to account for the possibility that states follow divergent long-run trajectories in both the outcome and in the likelihood of electing populist governors. This is a demanding specification: identification of β_1 and β_3 comes from within-state deviations from a state's own linear trend, net of common year shocks.

Standard errors are clustered at the state level throughout, allowing for arbitrary serial correlation and heteroskedasticity within states.

6 Results

Table 3 presents estimates of Equation (5.0.1) using repeal-adjusted civil-service reform as the institutional moderator. We build up the specification progressively. Column (1) includes the high-populism indicator, repeal-adjusted civil-service reform, their interaction, and state and year fixed effects. Column (2) adds state-specific linear trends. Column (3) further adds speech controls and governor party indicators. We focus the discussion on column (3), the most demanding specification, which identifies the effects from within-state deviations from each state's own linear trajectory, net of common year shocks, controlling for speech quantity, speech length, and governor party affiliation.

In the full specification, high-populism governors are associated with per-capita income roughly 3.8% lower in states where civil-service reform is not in force ($\hat{\beta}_1 = -0.038$, $p < 0.01$). The interaction with civil-service reform in force is positive, statistically significant, and larger in magnitude ($\hat{\beta}_3 = 0.049$, $p < 0.01$). The implied net association of high populism in states

[12]The additional controls: black population share, foreign-born population share, number of firms, and number of bankruptcies are sparsely observed in annual form. We construct filled state-year series by linear interpolation and extrapolation within state before taking logs; see Appendix H.5 for details.

Table 3: Populism, civil service reform, and log per capita income

	Civil service reform, accounting for repeals		
	(1)	(2)	(3)
High populism	-0.075^{***}	-0.037^{***}	-0.038^{***}
	(0.025)	(0.012)	(0.012)
Civil service reform in force	-0.010	0.005	0.005
	(0.023)	(0.015)	(0.015)
High populism × civil service reform in force	0.079^{**}	0.048^{***}	0.049^{***}
	(0.033)	(0.014)	(0.013)
Observations	3,861	3,861	3,861
States	50	50	50
Year FE	Yes	Yes	Yes
State FE	Yes	Yes	Yes
Controls	No	No	Yes
State trends	No	Yes	Yes
Within R^2	0.016	0.662	0.666

Notes: The dependent variable is ln(Per capita personal income). High populism is the binary indicator *Pop*, equal to one when the governor's mean LLM score exceeds the 90th percentile of the 1866-2023 state-year level distribution. Civil service reform is measured using `cs_ref_with_repeal`, which equals one when a civil service reform is in force after accounting for observed repeals through 2011. For post-2011 years, the 2011 reform status is carried forward. Controls include number of speeches, log token count, and party indicators. Standard errors clustered at the state level are reported in parentheses. One state-specific linear trend is omitted for collinearity. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

with civil-service reform in force is therefore close to zero and slightly positive: $-0.038+0.049 = 0.011$. The income penalty is concentrated in states without operative civil-service protections. The pattern is consistent across all three columns: the populism coefficient is negative and the interaction is positive and of comparable magnitude.

The analysis therefore focuses on institutional rules governing bureaucratic independence rather than administrative capacity. This choice is consistent with the cross-country evidence, where the attenuation pattern is strongest for measures of impartial and independent administration.

The U.S. analysis therefore centres on institutional rules governing bureaucratic independence rather than administrative capacity. This follows Vannoni et al. (2021) and Ash et al. (2022), who use civil-service reform to capture insulation from patronage, and aligns with the cross-country evidence in Section 3, where the attenuation pattern is strongest for measures of impartial and independent administration. The main channel through which bureaucracy dampens the economic effect of populism thus appears to be its independence rather than its scale. The distinction is intuitive: a large bureaucracy is not in itself a check on the executive. A populist governor who controls a sizeable administration may be better placed to implement their agenda than one with a small, under-resourced apparatus — so that under a spoils system a strong bureaucracy becomes a tool of the populist government rather than a constraint on it.

6.1 Interpretation

The state-level evidence points to a clear institutional moderation pattern: populist governors are associated with lower per-capita income when civil-service reform is not in force, but this penalty is attenuated—and in the main specification more than offset—where repeal-adjusted civil-service reform remains in force. The results therefore point specifically to institutional rules that insulate bureaucrats from patronage and discretionary political control. This interpretation is consistent with Ash et al. (2022) and, as noted above, with the cross-country evidence in Section 3. The state-level analysis further corroborates this pattern using within-country variation and more granular institutional measures.

We do not present bureaucratic quantity results because state-government headcount is unavailable for the pre-war period, and the interaction results are not precisely estimated across specifications.

6.2 Robustness

We probe the sensitivity of these results along several dimensions.

Alternative populism thresholds. Table 8 re-estimates the full specification using alternative definitions of populism while keeping the repeal-adjusted civil-service measure. The baseline indicator uses the 90th percentile. As robustness checks, we use a looser 75th-percentile indicator, a sign-based indicator equal to one whenever the governor's mean LLM score is positive, and the continuous governor-level mean LLM score. Across these alternatives, the populism coefficient remains negative and the interaction with civil-service reform in force remains positive. This suggests that the main result is not an artefact of the 90th-percentile cutoff.

Coefficient stability. Although we use a very demanding specification with both time and year fixed effects as well as state specific trends, unobservable variables may still pose an issue. Given that the phenomenon we are studying is far from being randomly assigned, standard causal inference techniques are not applicable. However, we can assess robustness to omitted variable bias using the framework of Oster (2019). For each interaction term, we compare the short regression (state and year fixed effects only) to the full specification and compute the degree of proportional selection on unobservables (δ^*) that would be required to drive the

coefficient to zero, setting $R_{\max} = \min\{1, 1.3 \times \hat{R}^2_{within}\}$. For the main $Pop\times$ civil-service-reform-in-force interaction, the degree of proportional selection on unobservables required to drive the coefficient to zero is $\delta^* = 5.35$. The bias-adjusted coefficient under equal selection remains positive at 0.0398, compared with the full-model estimate of 0.0490. These results indicate that unobservables would need to be more than five times as important as the full set of included controls: party, speech characteristics, and state-specific trends to explain away the interaction effects.

Additional controls. We also add a set of additional control variables. The underlying CSP controls are highly sparse in raw annual form, so we construct filled state-year series by interpolation and extrapolation within state before taking logs; Appendix H.5 documents this procedure and shows that the resulting logged controls are fully observed in the estimation sample. The additional controls are: (1) the Black population share, (2) the foreign-born population share, (3) the number of firms, and (4) the number of bankruptcies. We use these variables due to them predicting economic outcomes and populist takeovers. In the empirical specifications, we use the logged filled versions of these variables. Reassuringly, the interaction between high populism and civil-service reform in force remains positive and statistically significant after including these controls (Table 10). When we additionally include the first lag of log per-capita income, the interaction remains positive and statistically significant, although the magnitude declines, as expected in a dynamic specification.

Alternative populism measures. A potential concern is that the baseline Qwen score embeds model-specific idiosyncrasies. We therefore replace it with two dictionary-based alternatives and with GPT 5.5 populism scores. The dictionary measures use the updated lexicon of Gennaro et al. (2024), which expands and cleans the Pauwels (2011) vocabulary via WordNet, and the Pauwels (2011) dictionary itself. Both are constructed as non-compensatory TF–IDF scores requiring the simultaneous presence of anti-elite and people-centric stems; details appear in Appendix D.1. Across all specifications in Table 18, the interaction between dictionary-based populism and civil-service reform is positive, with the strongest results for continuous scores and 90th-percentile indicators. The dictionary estimates are noisier than the LLM estimates, which is unsurprising: keyword counts cannot capture the contextual, ideational features of populist discourse, in particular the Manichean opposition of a virtuous people to a corrupt elite, that holistic LLM grading is designed to detect (Tamaki et al., 2025).

We also re-score the full SOTS corpus using GPT 5.5 with the same holistic grading prompt and construct an analogous high-populism indicator. Although GPT 5.5 is not our preferred measure, since it tracks the human-coded validation sample less closely than Qwen and assigns small positive scores more often, the fixed effects results are strongly consistent with the baseline. In the full specification, high-populism governors are associated with lower per-capita income in states without civil-service reform in force ($\hat{\beta}_1 = -0.056$, $p < 0.01$), while the interaction with civil-service reform is positive and slightly larger in magnitude ($\hat{\beta}_3 = 0.062$, $p < 0.01$).

Civil service reform repeal. Because the main civil-service measure accounts for observed repeals through 2011 and carries forward the 2011 status thereafter, we examine whether the results depend on this coding assumption. First, we compare the main repeal-adjusted coding with the absorbing civil-service adoption indicator. Second, we restrict the sample to years before 2011, where the repeal information is directly observed and no post-2011 carry-forward assumption is required. Third, we restrict attention to 2000–2010, a period with near-complete speech coverage and directly observed repeal-adjusted coding. Results are reported in Appendix B, Table 9. The populism main effect remains negative and statistically significant at the 1% level across all three columns, while the interaction with civil-service reform is positive and significant at the 1% level in the full sample and the pre-2011 subsample, and at the 5%

level in the 2000–2010 window. The pattern is thus unchanged when institutional exposure is coded dynamically rather than as an absorbing state.

Lagged explanatory variables. A potential concern is reverse causality: economic downturns may increase demand for populist politicians, so that lower income causes populism rather than the reverse.

Several features of our design mitigate this concern. Civil-service reform is highly stable: it rarely changes within a state, and in most states adoption predates the regression period by decades, making it implausible that it responds to current economic conditions. State fixed effects and state-specific trends absorb the possibility that chronically poorer or economically declining states both elect more populist governors and have lower income independently. State of the State addresses are typically delivered in January or February, so the populism score is determined at the start of the calendar year, before most of that year's economic outcomes are realized. Column (2) of Table 11 replaces high populism, civil-service reform in force, their interaction, speech controls, and party indicators with their one-year lags. The lagged high-populism coefficient remains negative and statistically significant at the 1% level, $\hat{\beta}_1 = -0.0360$, and the lagged interaction with civil-service reform remains positive and statistically significant at the 1% level, $\hat{\beta}_3 = 0.0481$. The pattern is essentially unchanged from column (1), reducing the concern that the baseline estimates are driven by reverse causality.

Functional form and non-stationarity. A separate concern is that the choice of functional form: log per-capita personal income could drive the result, either because the log transformation matters mechanically or because the outcome variable is plausibly non-stationary, so the levels regression could in principle deliver spurious inference. Columns (3) and (4) of Table 11 address both concerns. Column (3) re-estimates the baseline using per-capita income in dollars: the interaction remains positive and statistically significant at the 5% level, $\hat{\beta}_3 = \$1{,}091$, fully offsetting the \$627 income shortfall associated with a high-populism governor in states without civil-service reform. Column (4) re-estimates the specification in first differences of log per-capita income, which are stationary by construction and therefore directly address the unit-root concern. The high-populism coefficient is $\hat{\beta}_1 = -0.0083$ ($p = 0.001$), implying a 0.83 percentage point lower annual income growth rate in states without civil-service reform, while the interaction is $\hat{\beta}_3 = +0.0084$ ($p = 0.004$), so the net effect under reform is near zero. The sign pattern is preserved across all four specifications.

Overall, we address three main threats to identification. First, potential measurement error in the populism score is addressed by replicating all results with dictionary-based alternatives (Gennaro et al., 2024; Pauwels, 2011) and GPT 5.5 populism scores. Second, reverse causality is addressed by lagging all key explanatory variables, exploiting the within-year timing of State of the State addresses, and noting the stability of civil-service reform over time. Third, omitted variable bias is addressed through a demanding specification with two-way fixed effects and state-specific linear trends, supplemented by controls and Oster (2019) coefficient stability bounds, which indicate that unobservables would need to be substantially more important than all included controls and the state specific trends to explain away the interaction effect. Finally, the consistency of our findings with the cross-country synthetic control evidence, and with a broader literature showing that institutional checks and balances attenuate the effects of leadership across a variety of settings (Jones and Olken, 2005; Clark et al., 2014; Ottinger and Voigtländer, 2025),lends further confidence that the moderating role of bureaucratic independence reflects a genuine causal mechanism rather than a statistical artefact.

7 Alternative empirical strategy: Difference in Difference

7.1 Estimator

To get further causal evidence for our hypothesis, we estimate the effect of populist gubernatorial control on state-level log per capita income using the heterogeneity-robust event-study estimator developed by de Chaisemartin and D'Haultfoeuille (2024).[13] This estimator is appropriate for our setting because the treatment is non-absorbing: states can both enter and later exit high-populism status. Standard two-way fixed-effects estimators are known to produce contaminated estimates in such designs when treatment effects are heterogeneous across units or over time (Goodman-Bacon, 2021). Furthermore, the event study version of the estimator allows us to gain some confidence in the causal interpretation of the results with the pre treatment placebos.

Here we shall briefly explain the setup, largely following the notation of de Chaisemartin and D'Haultfoeuille (2024); de Chaisemartin et al. (2025).

Let $g \in \{1, \dots, G\}$ index states and $t \in \{1, \dots, T\}$ index years.The outcome $Y_{g,t}$ is log per capita income; the binary treatment is $D_{g,t} = Pop_{g,t}$, equal to one when state g is classified as high-populism in year t. Our main specification uses an event window of eight years following each state's first populist takeover, spanning roughly two gubernatorial terms for both pre and post populist takeover.[14]

Target parameter. Let $Y_{g,t}(d_1, \dots, d_t)$ denote the potential outcome of state g in year t given the treatment history $(D_{g,1}, \dots, D_{g,t}) = (d_1, \dots, d_t)$. For each switching state g, let F_g denote the first year in which $D_{g,F_g} \neq D_{g,1}$, and let

$$T_g = \max_{g' : D_{g',1} = D_{g,1}} F_{g'} - 1$$

denote the last year in which there exists at least one state with the same period-one treatment as g whose treatment has not yet changed. The *actual-versus-status-quo* effect for state g at horizon ℓ is

$$\delta_{g,\ell} = \mathbb{E}\Big[Y_{g,F_g-1+\ell} - Y_{g,F_g-1+\ell}(D_{g,1}, \dots, D_{g,1}) \,\Big|\, \mathbf{D} \Big], \tag{7.1.1}$$

defined for g such that $F_g \leq T_g$ and $\ell \in \{1, \dots, T_g - F_g + 1\}$. The second term inside the expectation is the counterfactual outcome state g would have realized had it kept its period-one treatment status from period 1 through $F_g - 1 + \ell$. The parameter is conditional on the realized treatment matrix $\mathbf{D}$, following the design-based perspective of de Chaisemartin and D'Haultfoeuille (2024).

Switcher-specific estimator. Let $N_t^g = \#\{g' : D_{g',1} = D_{g,1},\ F_{g'} > t\}$ denote the number of states sharing g's baseline treatment whose treatment has not yet changed at year t. Lemma 1 of de Chaisemartin and D'Haultfoeuille (2024) establishes that

$$\begin{aligned} \widehat{\mathrm{DID}}_{g,\ell} = \big(Y_{g,\,F_g-1+\ell} - Y_{g,\,F_g-1}\big) - \\ - \frac{1}{N^g_{F_g-1+\ell}} \sum_{\substack{g' :\, D_{g',1} = D_{g,1} \\ F_{g'} > F_g - 1 + \ell}} \big(Y_{g',\,F_g-1+\ell} - Y_{g',\,F_g-1}\big). \end{aligned} \tag{7.1.2}$$

is unbiased for $\delta_{g,\ell}$ under the no-anticipation and parallel-trends assumptions stated below. Equation (7.1.2) compares g's outcome change from the year before its first treatment change to ℓ years later, against the simple average of the corresponding outcome changes among states sharing g's baseline treatment whose treatment has not yet changed at $F_g - 1 + \ell$.

[13] See also de Chaisemartin et al. (2025) for implementation details.

[14] Forty-eight of fifty states use four-year gubernatorial terms with term limits typically permitting two consecutive terms.

Aggregation. For each switching state, define

$$S_g = \begin{cases} +1 & \text{if } D_{g,F_g} > D_{g,1} \quad \text{(switcher-in)}, \\ -1 & \text{if } D_{g,F_g} < D_{g,1} \quad \text{(switcher-out)}, \end{cases}$$

following the sign convention of de Chaisemartin and D'Haultfoeuille (2024). In our binary setting, $S_g = +1$ states adopt populism and $S_g = -1$ states exit it. With $N_\ell = \#\{g : F_g - 1 + \ell \leq T_g\}$, the aggregate event-study coefficient is

$$\widehat{DID}_\ell = \frac{1}{N_\ell} \sum_{\substack{g: \\ F_g - 1 + \ell \leq T_g}} S_g \widehat{DID}_{g,\ell}. \tag{7.1.3}$$

The factor S_g ensures both adoptions and exits are expressed on a common"greater exposure to populism" scale. $\widehat{DID}_\ell$ is unbiased for the population non-normalized event-study effect δ_ℓ, defined as a weighted average of the $\delta_{g,\ell}$ across switchers (de Chaisemartin et al., 2025, Section 2.2).

Average total effect. Because Restriction 3 of de Chaisemartin and D'Haultfoeuille (2024) fails in our setting (states both adopt and exit populism), we follow the implementation in de Chaisemartin et al. (2025), which computes the average total effect per unit of treatment separately for switchers-in and switchers-out before combining them.

Let $N_\ell^+ = \#\{g : S_g = +1,\, F_g - 1 + \ell \leq T_g\}$ and $N_\ell^- = \#\{g : S_g = -1,\, F_g - 1 + \ell \leq T_g\}$ denote the number of switcher-in and switcher-out states for which $\widehat{DID}_{g,\ell}$ can be estimated at horizon ℓ, with $L^u = \max_{g:S_g=+1}(T_g - F_g + 1)$ and $L^a = \max_{g:S_g=-1}(T_g - F_g + 1)$ the corresponding maximal horizons. The horizon-specific event-study effects for switchers-in and switchers-out are

$$\widehat{DID}_{+,\ell} = \sum_{\substack{g:S_g=+1 \\ F_g-1+\ell \leq T_g}} \frac{N_{g,F_g-1+\ell}}{N_\ell^+} \widehat{DID}_{g,\ell}, \qquad \widehat{DID}_{-,\ell} = \sum_{\substack{g:S_g=-1 \\ F_g-1+\ell \leq T_g}} \frac{N_{g,F_g-1+\ell}}{N_\ell^-} \big(- \widehat{DID}_{g,\ell} \big).$$

The minus sign in $\widehat{DID}_{-,\ell}$ ensures that exits from treatment are expressed on the same "greater exposure to populism" scale as entries.

For switchers-in, define the sample-share weights and average dose change at horizon ℓ:

$$w_{+,\ell} = \frac{N_\ell^+}{\sum_{\ell'=1}^{L^u} N_{\ell'}^+}, \qquad \delta^D_{+,\ell} = \sum_{\substack{g:S_g=+1 \\ F_g-1+\ell \leq T_g}} \frac{N_{g,F_g-1+\ell}}{N_\ell^+} \big(D_{g,F_g-1+\ell} - D_{g,1} \big).$$

The switcher-in average total effect is

$$\widehat{\delta}_+ = \frac{\sum_{\ell=1}^{L^u} w_{+,\ell}\, \widehat{DID}_{+,\ell}}{\sum_{\ell=1}^{L^u} w_{+,\ell}\, \delta^D_{+,\ell}}.$$

For switchers-out, define analogously

$$w_{-,\ell} = \frac{N_\ell^-}{\sum_{\ell'=1}^{L^a} N_{\ell'}^-}, \qquad \delta^D_{-,\ell} = \sum_{\substack{g:S_g=-1 \\ F_g-1+\ell \leq T_g}} \frac{N_{g,F_g-1+\ell}}{N_\ell^-} \big(D_{g,1} - D_{g,F_g-1+\ell} \big).$$

Note the reversed sign in $\delta^D_{-,\ell}$: switchers-out experience treatment decreases, so $D_{g,1} - D_{g,F_g-1+\ell} \geq 0$, placing $\delta^D_{-,\ell}$ on the same scale as $\delta^D_{+,\ell}$. The switcher-out average total effect is

$$\widehat{\delta}_- = \frac{\sum_{\ell=1}^{L^a} w_{-,\ell}\, \widehat{DID}_{-,\ell}}{\sum_{\ell=1}^{L^a} w_{-,\ell}\, \delta^D_{-,\ell}}.$$

The two are combined into

$$\widehat{\delta} = w_+ \, \widehat{\delta}_+ + (1 - w_+) \, \widehat{\delta}_-,$$

where

$$w_+ = \frac{\left(\sum_{\ell=1}^{L^u} w_{+,\ell} \, \delta^D_{+,\ell}\right) \cdot \sum_{\ell'=1}^{L^u} N^+_{\ell'}}{\left(\sum_{\ell=1}^{L^u} w_{+,\ell} \, \delta^D_{+,\ell}\right) \cdot \sum_{\ell'=1}^{L^u} N^+_{\ell'} + \left(\sum_{\ell=1}^{L^a} w_{-,\ell} \, \delta^D_{-,\ell}\right) \cdot \sum_{\ell'=1}^{L^a} N^-_{\ell'}}$$

is the share of cumulative absolute treatment-dose changes accounted for by switchers-in. $\widehat{\delta}$ is unbiased for the average total effect per unit of treatment defined in Section 3.3 of de Chaisemartin and D'Haultfoeuille (2024), extended to designs where Restriction 3 fails. It is the central magnitude of interest: the average effect of being exposed to populism, expressed per unit of treatment, across the post-treatment window.

Normalization. Our main event-study results report the *normalized* effects $\widehat{DID}^n_\ell$ rather than $\widehat{DID}_\ell$. For each switcher g and horizon ℓ, let

$$\delta^D_{g,\ell} = \sum_{k=0}^{\ell-1} \left(D_{g,F_g+k} - D_{g,1}\right)$$

denote the cumulative difference between g's actual and status-quo treatment doses from F_g to $F_g - 1 + \ell$. Aggregating absolute dose changes across switchers,

$$\delta^D_\ell = \sum_{\substack{g \\ F_g - 1 + \ell \leq T_g}} \frac{N_{g,F_g-1+\ell}}{N^+_\ell + N^-_\ell} \left|\delta^D_{g,\ell}\right|, \qquad \widehat{DID}^n_\ell = \frac{\widehat{DID}_\ell}{\delta^D_\ell}.$$

de Chaisemartin and D'Haultfoeuille (2024) (Lemma 2) show that $\widehat{DID}^n_\ell$ is unbiased for the normalized event-study effect δ^n_ℓ, which measures a weighted average of the effects of the contemporaneous treatment and its first $\ell - 1$ lags on the period-$F_g - 1 + \ell$ outcome, with non-negative weights that sum to one and can themselves be estimated. Normalizing thus converts the horizon-ℓ event-study coefficient from an effect of "cumulative exposure to a weakly higher treatment dose for ℓ periods", whose magnitude depends on the realized treatment path, into a per-unit-of-treatment object with a transparent lag-effect interpretation. We do not apply an analogous normalization to the average total effect $\widehat{\delta}$: as a ratio of cumulative effects to cumulative dose changes, $\widehat{\delta}$ is already expressed per unit of treatment by construction (see Section 3.3 of de Chaisemartin and D'Haultfoeuille, 2024).

Identification. Equations (7.1.2)–(7.1.3) identify their population analogs under two assumptions (de Chaisemartin and D'Haultfoeuille, 2024):

- *No anticipation*: a state's outcome at t does not depend on its treatments after t.
- *Parallel trends for status-quo outcomes*: among states with the same period-one populism status, the expected evolution of log per capita income absent any change in populism is the same. This is weaker than the unconditional parallel trends assumption invoked by standard TWFE specifications

The placebo tests described below provide the most direct empirical check.

Placebos. The estimator computes pre-treatment placebo coefficients $\widehat{DID}_{\text{placebo},\ell}$, defined symmetrically to $\widehat{DID}_\ell$ but replacing $Y_{g,F_g-1+\ell} - Y_{g,F_g-1}$ with $Y_{g,F_g-1-\ell} - Y_{g,F_g-1}$. Intuitively, the placebos compare the outcome evolution of switchers included in $\widehat{DID}_\ell$ to that of their controls over the ℓ years *before* the switcher's first treatment change, rather than the ℓ years after. They thus assess whether switchers and controls were on parallel status-quo trajectories over the same length of pre-period that parallel trends must hold for $\widehat{DID}_\ell$ to be unbiased.

Heterogeneity by checks and balances. To test the mediation hypothesis that civil-service-based checks attenuate populism's economic effects, we partition switchers by their pre-takeover civil service regime, $X_g \in \{0, 1\}$, defined as the value of state g's civil service indicator in the year immediately preceding its first populist takeover. We re-estimate $\widehat{DID}_\ell$ separately within each subgroup using the package's `by` option and test whether the trajectories differ.[15] We construct two operationalizations of X_g. Our main specification uses civil service reform as in the previous part of the paper.

Sample restrictions. We restrict the panel to 1929 onwards, the first year for which per capita income is consistently observed across states. Missing populism treatment values, arising when no gubernatorial speech material is available to classify the governor, are handled by the package's default ("liberal") imputation rule. Two features of this default matter for our setting. First, the package redefines $D_{g,1}$ as state g's treatment at the first year its treatment is observed, rather than mechanically using the first calendar year of the panel, and defines F_g as the first year g is observed with a treatment different from that initial observation. Second, missing treatments between the first observation and the first switch are imputed with g's redefined baseline treatment; missing treatments after the first switch are imputed with D_{g,F_g}. Outcomes before the first observation of treatment are set to missing, so those state-years drop out of estimation. As a robustness check, we use the `drop_if_d_miss_before_first_switch` option, which drops groups with missing treatment before their first switch rather than applying the package's liberal pre-switch imputation rules.[16]

7.2 Results

Defining the civil service moderator. The de Chaisemartin and D'Haultfoeuille (2024) estimator with the `by()` option requires a time-invariant, state-level moderator: each state is permanently assigned to one subgroup, and the estimation is run separately within each subgroup. This requirement is in tension with the substantive concept we want to capture, the strength of civil service-based checks on executive discretion at the moment of populist takeover, because civil service regimes vary over time within a state and the relevant year differs across switchers. We considered several operationalizations and discuss why each was unsatisfactory before settling on our preferred specification.

The most natural definition codes each state by its civil service status in the year before its first populist takeover, B_{g,F_g-1}. This has two related problems. First, the comparison logic of the `by()` option requires that controls share the switcher's moderator value, but a state's status in year $F_g - 1$ may not match its status in the actual comparison years post-F_g. A state coded $B = 1$ at takeover may have repealed reform shortly afterward, in which case its post-takeover trajectory is not informative about how "B-strong" environments shape populist effects. Second, control states whose own $F_{g'}$ falls much later may share the switcher's B value in $F_g - 1$ but have an entirely different civil service regime in the years that matter for the comparison. The single-year snapshot is descriptively incomplete.

A second candidate codes states by lifetime civil service tradition, for example, $B = 1$ if civil service reform was in force for at least the median share of years in 1929–2023. This sidesteps the single-year fragility but introduces an era-mixing problem: a state that was strongly CB-aligned across most of the modern era may have experienced its populist takeover in a window

[15] We do not use the package's `predict_het` option, which builds on de Chaisemartin and D'Haultfoeuille (2024)'s Lemma 6 to estimate $\hat{\beta}^{het}_\ell$, the coefficient on X_g in a regression of group-level event-study estimates on X_g saturated with indicators for $F_g \times D_{g,1} \times S_g$ cohorts. Because state-level populist takeovers occur in nearly unique years across our switchers, most cohort cells contain a single state, rendering the cohort-saturated regression numerically degenerate.

[16] See Appendix B of de Chaisemartin et al. (2025) for the full set of rules, including the treatment of states whose populism status never changes over the panel.

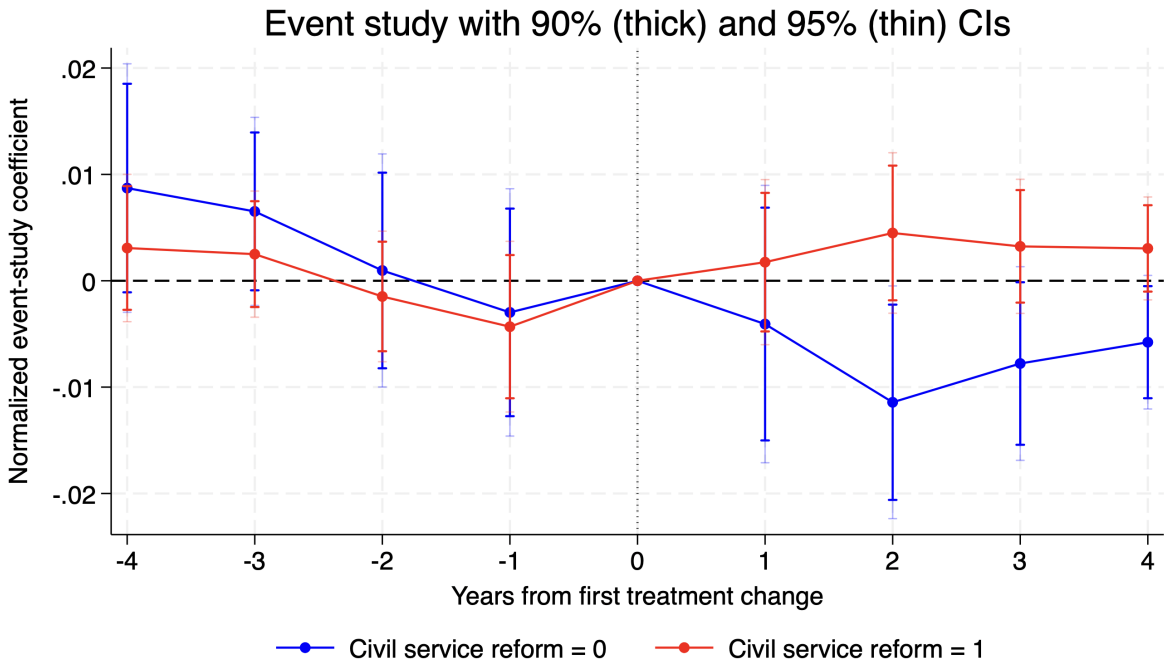


Figure 5: Event-study estimates of populism's effect on log per capita income, by pre-takeover civil service reform status. Solid lines plot $\widehat{DID}_\ell^n$ for $\ell = -4, \ldots, 4$, separately for states without civil-service reform in force throughout the 2010–2023 estimation window (`B_modern` $= 0$) and states with civil service reform in force throughout (`B_modern` $= 1$). Solid vertical bars show cluster-robust 90% confidence intervals; faint vertical bars show cluster-robust 95% confidence intervals.

when reform was not yet adopted. Concretely, suppose a state has $B = 1$ under this operationalization (reform in force for most of the panel) but its populist takeover occurred in 1935, before reform was even enacted. The substantively correct counterfactual for that takeover is a $B = 0$ environment, but the lifetime measure assigns $B = 1$ and the comparison is muddled.

To address both problems, we restrict our attention to the post-2010 panel, exploiting two features of the data. First, the wave of civil service reform repeals concentrated in the 1996–2005 period (15 states repealed in this window) had concluded by 2005: $B_{g,t}$ is a state-level constant for every state from 2005 through 2023. Second, restricting the sample to year ≥ 2010 allows a five-year buffer for institutional consolidation following the repeal wave: a reform repealed in 2003 has had seven years to take its current form by 2010, by which point the relevant bureaucratic culture, personnel, and enforcement capacity reflect the post-repeal regime rather than residual structures from the pre-repeal era. We define `B_modern` as the (constant-by-construction) value of $B_{g,t}$ within the 2010–2023 window: `B_modern` $= 1$ for the 34 states with civil service reform in force throughout the period, and `B_modern` $= 0$ for the 16 states without reform in force (comprising the 15 repealers plus Texas, which never adopted reform).This definition has the advantage that the moderator describes the actual contemporaneous institutional environment in every year of the estimation sample, not a snapshot or aggregate.

One feature of the post-2010 sample is worth noting: the missing-treatment problem is essentially absent. Only one governorship-year (Jeff Colyer's 2018–2019 term in Kansas) has missing populism data within a switcher's pre- or post-takeover window, and Colyer was a transitional governor with no State of the State address available for classification.

Table 4: Event-study estimates: high populism's effect on log per-capita income, by modern civil-service regime

	`B_modern` = 0 (no civil-service reform in force)			`B_modern` = 1 (civil-service reform in force)		
	Estimate	SE	N/S	Estimate	SE	N/S
Pre-treatment placebos						
Placebo_4	0.009	0.006	35/5	0.003	0.004	72/11
Placebo_3	0.007	0.005	38/5	0.003	0.003	77/11
Placebo_2	0.001	0.006	39/5	−0.001	0.003	82/11
Placebo_1	−0.003	0.006	42/5	−0.004	0.004	84/11
Joint nullity test (placebos):		$p = 0.214$			$p = 0.276$	
Post-treatment effects						
Effect_1	−0.004	0.007	57/8	0.002	0.004	114/13
Effect_2	-0.011^{**}	0.006	54/8	0.004	0.004	112/13
Effect_3	-0.008^{*}	0.005	53/8	0.003	0.003	107/13
Effect_4	-0.006^{*}	0.003	50/8	0.003	0.002	102/13
Joint nullity test (effects):		$p = 0.012$			$p = 0.533$	
Av_tot_eff	-0.018^{*}	0.010	148/32	0.008	0.007	327/52
Avg. exposure (years)		2.50			2.54	

Notes: Estimates from `did_multiplegt_dyn` (de Chaisemartin et al., 2025) on the panel of U.S. states from 2010 onwards. The treatment is a binary indicator for high populism, *Pop*, equal to one when the governor's mean LLM populism score exceeds the 90th percentile of the state-year distribution of governor-mean LLM scores. The moderator `B_modern` separates states by their modern civil-service regime. Standard errors are clustered at the state level. The specification uses the `normalized`, `same_switchers`, and `same_switchers_pl` options to keep the composition of switchers fixed across event-study and placebo horizons. In the event-study and placebo rows, N/S reports the number of observations and the number of switchers contributing to each horizon. The number of switchers can differ between placebo and post-treatment rows because some states have a full four years of post-treatment data within the 2010–2023 window but fewer than four pre-treatment years. The `same_switchers` and `same_switchers_pl` options hold composition fixed within the post-treatment and placebo windows separately. In the Av_tot_eff row, N/S reports the number of observations and the number of switcher-periods. Av_tot_eff is the average cumulative total effect per unit of treatment, expressed in log points; "Avg. exposure" reports the average number of post-treatment periods over which the cumulative effect is computed. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$; based on cluster-robust standard errors.

Table 4 reports normalized event-study estimates of populism's effect on log per capita income for the 2010–2023 sample, separately for states without civil service reform in force during the modern era (`B_modern` = 0, $n = 8$ switchers) and states with reform in force throughout (`B_modern` = 1, $n = 13$ switchers).[17] Figure 5 plots the trajectories. The estimates are consistent with the main fixed-effects results. In states without civil-service reform in force, log per-capita income falls relative to control states across all four post-treatment horizons. The effects at horizons 2, 3, and 4 are individually significant at the 10% level (horizon 2 is also significant at 5%), and the joint test rejects the null of zero post-treatment effects ($p = 0.012$). The average total effect is −0.018, implying a decline of roughly 1.8 log points over the post-treatment window. By contrast, in states with civil-service reform in force, the post-treatment coefficients are small, positive, and statistically indistinguishable from zero; the joint post-treatment test does not reject ($p = 0.533$), and the average total effect is 0.008. This asymmetry is consistent with the interpretation that civil-service protections attenuate the economic costs of high-populism governance.

[17] We restrict to the post-2010 sample so that all state-level civil service regimes are time-invariant within the estimation window: the wave of reform repeals concentrated in the 1996–2005 period had concluded by 2005, and we allow a five-year buffer for institutional consolidation before beginning the analysis.

Pre-treatment placebos. The placebo evidence is reassuring. In both subgroups, the placebo estimates are small in magnitude, never exceeding 0.009 in absolute value, and none is individually significant. The joint placebo tests also fail to reject the null of no pre-treatment effects in either subgroup ($p = 0.214$ for `B_modern` $= 0$; $p = 0.276$ for `B_modern` $= 1$). Thus, the four-year pre-treatment window shows no evidence of differential trends before the switch into high-populism governance. This strengthens the interpretation that the post-treatment divergence is not simply capturing pre-existing income dynamics.

Post-treatment effects. The post-treatment estimates reveal a sharp difference by civil service regime. In states without modern civil service reform, high-populism governance is followed by a consistent decline in log per capita income. The estimated effect is negative at every post-treatment horizon. While the horizon-1 effect is small and not individually significant, the effects at horizons 2, 3, and 4 are statistically significant: -0.011, -0.008, and -0.006 log points, respectively. The joint post-treatment test rejects the null of no effect ($p = 0.012$). The average total effect is -0.018 log points, significant at the 10% level, implying an income decline of roughly 1.8% over an average exposure window of 2.5 years.

In contrast, states with modern civil service reform show no comparable decline. The post-treatment estimates are small, positive, and statistically indistinguishable from zero at every horizon. The joint test does not reject the null of no post-treatment effects ($p = 0.533$), and the average total effect is close to zero at 0.008 log points.

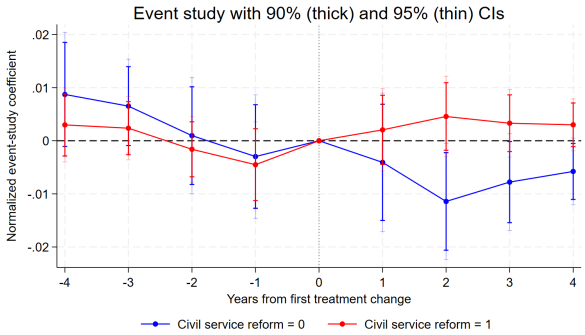
Heterogeneity. This divergence is the key event-study result. The negative effect of high-populism governance appears only in states without modern civil service reform. Where civil service reform is absent, the estimates are negative at all horizons, jointly significant, and economically meaningful. Where reform is in force, the estimates are near zero and statistically insignificant throughout. The pattern therefore provides strong corroborative evidence for the paper's central claim: bureaucratic institutions moderate the economic consequences of populist governance. In weak civil service environments, high-populism governors are followed by lower per capita income; in modern civil service environments, this effect is effectively muted.

We caveat these results by pointing out that the sample is small: 8 switchers in the no-reform subgroup and 13 in the reform subgroup within the 2010–2023 window.

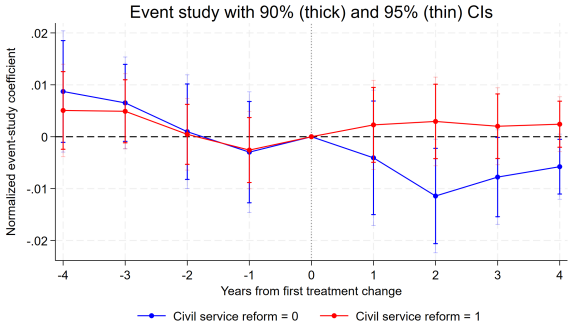
7.3 Robustness

In this section we report four robustness checks, summarised in Table 5 and reported in detail in Appendix Tables 12-15. These checks use, respectively, a strict-missing rule for treatment imputation, a switchers-in-only specification, an all-switchers specification without the same-switcher restriction, and a level-income specification. Figure 6 plots the corresponding event-study estimates. Across these specifications, the same substantive pattern emerges: post-treatment income declines are concentrated in states without modern civil service reform, while states with reform in force show no comparable decline.

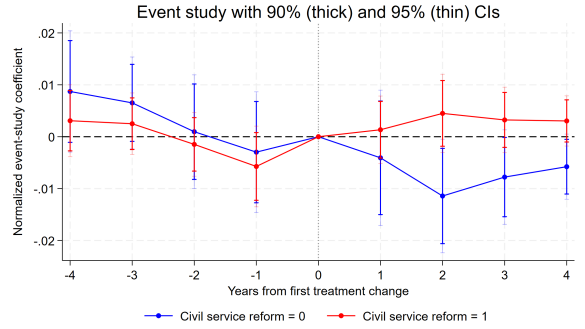
Strict missing values. The first robustness check applies a stricter rule for missing treatment values by using the `drop_if_d_miss_before_first_switch` option. This drops groups with missing treatment status before their first switch, rather than relying on the package's default liberal imputation rule. The results are nearly identical to the baseline specification. In states without modern civil service reform, the post-treatment effects remain negative at every horizon, with statistically significant estimates from horizon 2 onward and a joint post-treatment test that rejects the null ($p = 0.012$). In states with reform in force, the estimates remain small and statistically insignificant, and the joint post-treatment test does not reject ($p = 0.588$).



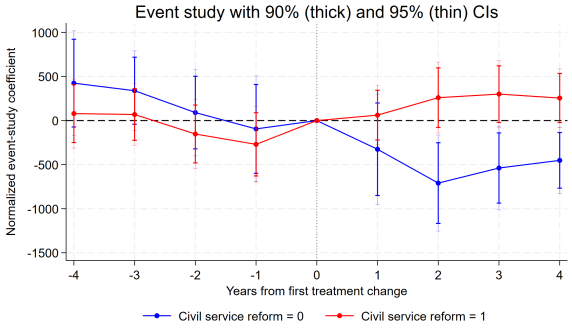
(a) Strict-missing rule



(b) Switchers into populism only



(c) All switchers (no same-switcher restriction)



(d) Per-capita income in levels

Figure 6: Robustness of the dynamic DiD event-study estimates. Each panel reports normalized event-study coefficients $\widehat{\mathrm{DID}}_\ell^n$ for $\ell = -4, \ldots, 4$, by modern civil-service reform status, under an alternative specification: (a) drops state-years with missing treatment before the first switch (`drop_if_d_miss_before_first_switch` option of `did_multiplegt_dyn`); (b) restricts attention to switches into high populism only (`switchers(in)` option); (c) drops the `same_switchers` and `same_switchers_pl` restrictions so the set of switchers may vary across horizons; (d) uses per-capita income in levels rather than logs as the outcome. Solid (faint) vertical bars show cluster-robust 90% (95%) confidence intervals. Across all four specifications, the negative post-takeover trajectory is concentrated in states without civil-service reform in force; no comparable decline appears in states with reform in force.

Table 5: Summary of dynamic DiD robustness checks: high populism's effect on per-capita income, by modern civil-service regime

	Joint p (placebos)		Joint p (effects)		Av. total effect	
Specification	`B_modern`=0	`B_modern`=1	`B_modern`=0	`B_modern`=1	`B_modern`=0	`B_modern`=1
Baseline	0.214	0.276	0.012	0.533	-0.018^{*}	0.008
Strict missing	0.214	0.309	0.012	0.588	-0.018^{*}	0.008
Switchers in only	0.214	0.036	0.012	0.849	-0.018^{*}	0.006
All switchers	0.214	0.112	0.012	0.548	-0.018^{*}	0.008
Levels ($, not logs)	0.176	0.204	0.005	0.555	$-1{,}292^{**}$	635

Notes: Each row reports the headline statistics from one DiD specification estimated with `did_multiplegt_dyn` (de Chaisemartin and D'Haultfoeuille, 2024; de Chaisemartin et al., 2025). "`B_modern`=0" denotes states without civil-service reform in force during 2010–2023 (8 switchers); "`B_modern`=1" denotes states with civil-service reform in force throughout this period (13 switchers in the baseline; the count varies slightly across specifications). Joint p-values are from the package's F-tests of nullity for the four pre-treatment placebos and the four post-treatment horizons and are reported directly. The Av. total effect column reports $\widehat{\delta}$ in log points except in the levels row, which reports dollars per unit of treatment. Detailed per-horizon tables are reported in Tables 12–15 in the Appendix. Asterisks on Av. total effect estimates indicate $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$; based on cluster-robust standard errors.

The placebo tests do not reject in either subgroup. Thus, the main result is not driven by the treatment imputation rule.

Switchers into high populism only. The second robustness check restricts attention to switches into high populism only, using the `switchers(in)` option. This addresses the possibility that the baseline result is partly driven by switches out of high populism. The core pattern remains unchanged. In states without modern civil service reform, the post-treatment estimates are negative at all horizons, statistically significant from horizon 2 onward, and jointly significant ($p = 0.012$). In states with reform in force, the post-treatment estimates are close to zero and jointly insignificant ($p = 0.849$). The main caveat is that the reform subgroup fails the joint placebo test in this specification ($p = 0.036$). However, no individual coefficient is significant at 10% and the coefficients are extremely small.

All switchers without the same-switcher restriction. The third robustness check removes the `same_switchers` and `same_switchers_pl` restrictions, allowing the set of switchers to vary across event-study and placebo horizons. This increases the use of available event-time observations but changes the composition of switchers across horizons. The results are again very similar to the baseline. In states without modern civil service reform, the post-treatment effects remain negative at all horizons and the joint post-treatment test rejects the null ($p = 0.012$). In states with reform in force, the estimates remain small and statistically indistinguishable from zero, with no joint rejection ($p = 0.548$). The placebo tests remain non-rejected in both subgroups. The conclusion is therefore not sensitive to holding the switcher composition fixed across horizons.

Per capita income in levels. The fourth robustness check replaces log per capita income with per capita income in levels. This is a useful functional-form check rather than a purely mechanical rescaling. As Roth and Sant'Anna (2023) emphasize, parallel trends is not necessarily invariant to monotonic transformations of the outcome. Equivalently, parallel trends in logs and parallel trends in levels are distinct identifying restrictions: the former identifies effects in relative terms, while the latter identifies effects in the original units of the outcome. Hence, we show that estimating the design under alternative outcome transformations helps assess whether the substantive conclusion does not depend on a particular functional-form assumption.

The level specification supports the baseline log-income result. In states without modern civil service reform, high-populism entry is followed by sizeable income declines: the post-treatment estimates are negative at all horizons, statistically significant from horizon 2 onward, and jointly significant ($p = 0.005$). The average total effect is approximately −\$1,292 per unit of treatment, with both the 90% and 95% confidence intervals excluding zero. In states with reform in force, the estimates are positive but statistically indistinguishable from zero, and the joint post-treatment test does not reject ($p = 0.555$). Thus, the level-income specification reinforces the main interpretation: income losses following high-populism entry are concentrated in states without modern civil service protections, and this pattern is not an artifact of imposing parallel trends in log income.

Alternative LLM populism measure. A final check replaces the Qwen-based treatment indicator with one constructed from an alternative large language model, GPT-5.5, using the same holistic-grading prompt and a 90th-percentile 1866–2023 state-year cutoff (0.25). We use GPT-5.5 only as a robustness check rather than as our main measure because it tracks the hand-coded human ratings less closely than Qwen on nearly every validation metric—lower exact agreement, higher mean absolute error, and a larger upward bias, as it assigns small positive scores to many speeches a human coder rates zero (see Appendix Table 25). A further difference from the Qwen run is that 34 speeches did not return parseable holistic reasoning under GPT-5.5 and were recovered with a score-only fallback that returns the scalar grade alone; results are robust to dropping these 34 (see Appendix D.2).

The directional pattern is nonetheless preserved: in states without modern civil-service reform the post-treatment coefficients are negative at every horizon and jointly significant ($p = 0.035$), with an average total effect of −0.010 log points, whereas in states with reform in force the effects are essentially zero (joint $p = 0.991$; average total effect 0.000).

Two caveats lead us to read this as weaker corroboration than the Qwen-based estimates rather than a clean replication. First, the magnitudes are smaller; none of the individual post-treatment coefficients reaches conventional significance, and the average total effect is not individually significant. Second, and more importantly, the pre-treatment placebo test rejects parallel trends in the no-reform subgroup (joint $p < 0.001$, with a significant placebo at $\ell = -2$), so the identifying assumption is not clearly satisfied for this measure. The switcher counts are also very small—4 switchers in the no-reform group and 8 in the reform group—which limits power and stability. We therefore treat the GPT-based DiD as directionally consistent with our main results while emphasizing that it does not meet the same pre-trend standard; full estimates are in Appendix Table 20.

8 Conclusion

This paper provides a range of evidence that an independent bureaucracy mediates the negative economic consequences of populist government. In the cross-country analysis, we replicate the findings of Funke et al. (2023) that populist takeovers are followed by meaningful declines in GDP per capita, and show evidence that this effect is attenuated in countries with stronger pre-existing bureaucratic institutions. Given the small number of treated episodes available for the split-sample analysis, we treat these results as suggestive: they establish a pattern consistent with the broader literature on populism and checks and balances and motivate the sharper within-country analysis pursued in the U.S. setting. The heterogeneity appears to be driven primarily by the rigorous-and-impartial-administration component of our composite, consistent with bureaucratic independence, rather than administrative scale or compensation alone. The attenuation effect peaks after 4 years but does not persist in the long term, suggesting that independent bureaucracies may delay or soften the initial output cost of populism rather than

eliminating it entirely, or that populist governments erode these checks and balances over time, as suggested by Kyriacou and Trivín (2025).

The U.S. state-level evidence sharpens this picture considerably. Exploiting within-state variation in both populist governance and institutional design across an unbalanced panel starting from 1929, we find that populist governors are associated with approximately 3.8% lower per-capita income in states where civil-service reform is not in force. This penalty is more than offset in states where civil-service reform remains in force after accounting for observed repeals. These findings survive a demanding set of robustness checks: alternative populism definitions, replacement of the LLM-based measure with dictionary-based alternatives, alternative civil-service coding and sample restrictions, additional demographic and economic controls, lagged specifications, and coefficient-stability bounds following Oster (2019).

As a complementary empirical strategy, we also estimate a heterogeneity-robust dynamic DiD in the post-2010 sample. This exercise exploits switches into high-populism status and asks whether income changes after high-populism entry, separately by whether civil-service reform is in force. We restrict the DiD analysis to the post-2010 period because speech coverage is near-complete and civil-service status is stable within the event-study window. The results point in the same direction as the main panel estimates: high-populism entry is followed by income declines in states without civil-service reform in force, while no comparable decline is detected in states with reform in force. Because this design uses a shorter subsample, we interpret it as timing-based corroborating evidence rather than as the paper's primary estimate.

Taken together, this evidence suggests that when a populist leader takes office, an independent bureaucracy partially attenuates the resulting output loss by constraining the implementation of economically distortionary policies. This finding bridges the gap between different strands of the political economy and political science literature that focus on the consequence of populist government and the tension between populism and bureaucracy. Moreover, this has a direct policy implication. Efforts to professionalize and insulate the civil service are not merely good governance in normal times: they provide structural insurance against the economic costs of populist governance. Paradoxically, the institutional "swamp" that populists promise to drain is precisely the feature that limits the damage they would otherwise cause.

References

Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, June 2021. doi: 10.1257/jel.20191450.

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132, March 2003. doi: 10.1257/000282803321455188.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 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, 2010. doi: 10.1198/jasa.2009.ap08746.

Elliott Ash, Massimo Morelli, and Matia Vannoni. Divided government, delegation, and civil service reform. *Political Science Research and Methods*, 10(1):82–96, 2022. doi: 10.1017/psrm.2020.51. URL https://www.cambridge.org/core/journals/political-science-research-and-methods/article/divided-government-delegation-and-civil-service-reform/BB0093CF1DE04CCF4F84BA387BFF2F33.

Michael W. Bauer and Stefan Becker. Democratic backsliding, populism, and public administration. *Perspectives on Public Management and Governance*, 3(1):19–31, 2020. doi: 10.1093/ppmgov/gvz026.

Luca Bellodi, Massimo Morelli, Antonio Nicolò, and Paolo Roberti. The shift to commitment politics and populism: Theory and evidence. CEPR Discussion Paper 18338, CEPR, July 2023.

Luca Bellodi, Massimo Morelli, and Matia Vannoni. A costly commitment: Populism, economic performance, and the quality of bureaucracy. *American Journal of Political Science*, 68(1):193–209, 2024. doi: 10.1111/ajps.12782.

Luca Bellodi, Massimo Morelli, Jörg L Spenkuch, Edoardo Teso, Matia Vannoni, and Guo Xu. Personnel is policy: Delegation and political misalignment in the rulemaking process. Technical report, National Bureau of Economic Research, 2026.

Timothy Besley, Robin Burgess, Adnan Khan, and Guo Xu. Bureaucracy and development. *Annual Review of Economics*, 14:397–424, 2022. doi: 10.1146/annurev-economics-080521-011950.

Fred Bever. Governor imposes moratorium on new state wind development permits. *Maine Public*, 2018.

Rod R. Blagojevich. State of the state address, 2003.

Rod R. Blagojevich. State of the state address, 2008.

John G. Bullock. Education and attitudes toward redistribution in the united states. *British Journal of Political Science*, 2020.

Timothy R. Clark et al. When do leaders matter? Ownership, governance and the influence of CEOs. *Leadership Quarterly*, 25(2):358–372, 2014. doi: 10.1016/j.leaqua.2013.09.004.

Clément de Chaisemartin and Xavier D'Haultfoeuille. Difference-in-differences estimators of intertemporal treatment effects. *Review of Economics and Statistics*, pages 1–45, 2024. doi: 10.1162/rest_a_01414. URL https://doi.org/10.1162/rest_a_01414.

Clément de Chaisemartin, Bingxue Li, Diego Ciccia, Mélitine Malézieux, Doulo Sow, David Arboleda, Romain Angotti, Xavier d'Haultfœuille, Felix Knau, Henri Fabre, and Anzony Quispe Rojas. Using did_multiplegt_dyn to estimate event-study effects in complex designs: Overview, and four examples based on real datasets. Ssrn working paper, SSRN, July 2025. URL https://ssrn.com/abstract=5337463. Posted July 22, 2025; last revised October 8, 2025.

Rudiger Dornbusch and Sebastian Edwards. Macroeconomic populism. *Journal of Development Economics*, 32(2):247–277, 1990. doi: 10.1016/0304-3878(90)90038-D.

Efficiency Maine Trust. FY2018 Annual Report, 2018. Fiscal year July 1, 2017–June 30, 2018; revised November 30, 2018.

Manuel Funke, Moritz Schularick, and Christoph Trebesch. Populist leaders and the economy. *American Economic Review*, 113(12):3242–3288, 2023. doi: 10.1257/aer.20202045.

Gloria Gennaro, Giampaolo Lecce, and Massimo Morelli. Intertemporal evidence on the strategy of populism in the united states. Working paper, May 6, 2024, 2024.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021. doi: 10.1016/j.jeconom.2021.03.014.

Matt Grossmann, Marty Jordan, and Joshua McCrain. The correlates of state policy and the structure of state panel data. *State Politics & Policy Quarterly*, 2021. doi: 10.1017/spq.2021.17.

Ari Hoogenboom. The Pendleton act and the civil service. *The American Historical Review*, 64 (2):301–318, 1959. doi: 10.1086/ahr/64.2.301.

Illinois Civil Service Commission. About the civil service commission, 2026.

Benjamin F. Jones and Benjamin A. Olken. Do leaders matter? National leadership and growth since World War II. *Quarterly Journal of Economics*, 120(3):835–864, 2005. doi: 10.1093/qje/120.3.835.

Jacob Kaplan. United States Governors 1775-2020, 2021. URL https://doi.org/10.3886/E102000V3. Version 3.

Andreas P. Kyriacou and Pedro Trivín. Populism and the rule of law: The importance of institutional legacies. *American Journal of Political Science*, 2025. doi: 10.1111/ajps.12935.

Paul LePage. State of the state address, 2013.

Paul LePage. State of the state address, 2014.

Paul LePage. State of the state address, 2016.

Paul LePage. State of the state address, 2017.

Paul LePage. State of the state address, 2018.

Nicolas E Magud and Antonio Spilimbergo. Economic and institutional consequences of populism. 2021.

Maine Legislature. Efficiency maine trust: Governance and board. Maine Revised Statutes, Title 35-A, Section 10103, 2012. The statute establishes the Trust as a public instrumentality governed by the independent Efficiency Maine Trust Board.

Kenneth J. Meier, Mallory Compton, John Polga-Hecimovich, Miyeon Song, and Cameron Wimpy. Bureaucracy and the failure of politics: Challenges to democratic governance. *Administration & Society*, 51(10):1576–1605, 2019. doi: 10.1177/0095399719874759.

Kevin Miller. Lepage blocks new wind energy projects, creates secretive commission to study impacts. *Portland Press Herald*, 2018.

Donald P. Moynihan, Pamela Herd, and Hope Harvey. Administrative burden: Learning, psychological, and compliance costs in citizen-state interactions. *Journal of Public Administration Research and Theory*, 25(1):43–69, 2015.

Cas Mudde. The populist zeitgeist. *Government and Opposition*, 39(4):541–563, 2004. doi: 10.1111/j.1477-7053.2004.00135.x.

Cas Mudde and Cristóbal Rovira Kaltwasser. Exclusionary vs. inclusionary populism: Comparing contemporary europe and latin america. *Government and opposition*, 48(2):147–174, 2013.

Jan-Werner Müller. *What Is Populism?* University of Pennsylvania Press, Philadelphia, 2016. ISBN 978-0-8122-4898-2. doi: 10.9783/9780812293784.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019. doi: 10.1080/07350015.2016.1227711.

Sebastian Ottinger and Nico Voigtländer. History's masters: The effect of European monarchs on state performance. *Econometrica*, 93(1):95–128, 2025. doi: 10.3982/ECTA20830.

Teun Pauwels. Measuring populism: A quantitative text analysis of party literature in belgium. *Journal of Elections, Public Opinion and Parties*, 21(1):97–119, 2011.

B. Guy Peters and Jon Pierre. Populism and public administration: Confronting the administrative state. *Administration & Society*, 51(10):1521–1545, 2019. doi: 10.1177/0095399719888960.

Jonathan Roth and Pedro H. C. Sant'Anna. When is parallel trends sensitive to functional form? *Econometrica*, 91(2):737–747, 2023. doi: 10.3982/ECTA19402.

Greg Sasso and Massimo Morelli. Bureaucrats under populism. *Journal of Public Economics*, 202:104497, 2021. doi: 10.1016/j.jpubeco.2021.104497.

Jay M. Shafritz, E. W. Russell, and Christopher P. Borick. *Introducing Public Administration*. Pearson, Boston, 8th edition, 2012. ISBN 978-0205855896.

Stephen Skowronek. *Building a New American State: The Expansion of National Administrative Capacities, 1877–1920*. Cambridge University Press, Cambridge, 1982. doi: 10.1017/CBO9780511665080.

State of Illinois. Personnel code, 20 ilcs 415, 2026.

STLPR. Two fired state workers reinstated. https://www.stlpr.org/other/2007-08-08/two-fired-state-workers-reinstated, August 2007. Accessed 18 June 2026.

Eduardo Ryô Tamaki, Yujin J. Jung, Julia Chatterley, Grant Mitchell, Semir Dzebo, Cristóbal Sandoval, Levente Littvay, and Kirk A. Hawkins. Populism meets ai: Advancing populism research with llms, 2025.

U.S. Census Bureau. Statistics of u.s. businesses: Susb totals for u.s. & states. Dataset / statistical table, 2012.

U.S. Congress, House. Committee on Post Office and Civil Service. Subcommittee on Manpower and Civil Service. History of civil service merit systems of the United States and selected foreign countries. Committee print, 94th congress, 2nd session, U.S. Government Printing Office, Washington, DC, 1976.

U.S. DOJ. Former illinois governor rod blagojevich, his brother, two former top aides, and two businessmen indicted. Press release, Federal Bureau of Investigation, Chicago Field Office, 2009.

V-Dem Institute. Varieties of democracy (V-Dem) dataset v16. Dataset, 2026. URL https://v-dem.net/data/the-v-dem-dataset/. Version 16.

Matia Vannoni, Elliott Ash, and Massimo Morelli. Measuring discretion and delegation in legislative texts: Methods and application to us states. *Political Analysis*, 29(1):43–57, 2021.

Kurt Weyland. Clarifying a contested concept: Populism in the study of Latin American politics. *Comparative Politics*, 34(1):1–22, 2001. doi: 10.2307/422412.

Justin Worland. Maine to test some welfare recipients for drugs. *TIME*, 2015.

A Descriptive tables

Table 6: Civil Service Reform and High Populism: Joint Distribution

	High populism (*Pop*)		
Civil service status	0 (Low)	1 (High)	Total
No reform / repealed	1,010	132	1,142
	(88.4%)	(11.6%)	
Reform in force	2,443	276	2,719
	(89.9%)	(10.1%)	
Total	3,453	408	3,861

Notes: Each cell reports the count of state-year observations in the baseline regression sample, with row percentages in parentheses. *Pop* equals one when the governor's mean LLM score exceeds the 90th percentile of the 1866-2023 state-year distribution. Civil service reform is measured using `cs_ref_with_repeal`, which equals one when a civil service reform is in force after accounting for observed repeals through 2011. For post-2011 years, the 2011 reform status is carried forward.

Table 7: Distribution of LLM Populism Scores: Speech Level

	N	Mean	SD	Min	Median	P75	P90	P95	Max
LLM score	3,355	0.024	0.099	0.000	0.000	0.000	0.100	0.200	1.400
Word count	3,355	5,897	4,676	201	4,649	6,662	10,076	14,181	67,912

Notes: Statistics are computed at the speech level across the full SOTS qwen scoring corpus of 3,355 speeches. Of these, 2,963 receive a score of zero and 392 (11.7%) receive a positive score. Word count is computed for all 3,355 speeches from the raw processed speech text before aggregation to the governor or state-year level.

B Additional USA robustness tables

Table 8: Populism, civil service reform, and log per capita income: alternative populism measures

	Civil service reform, accounting for repeals		
	$p > 75$	$p > 0$	Continuous score
	(1)	(2)	(3)
Populism	-0.041^{***}	-0.032^{***}	-0.118^{**}
	(0.011)	(0.011)	(0.058)
Civil service reform in force	0.000	0.001	0.006
	(0.014)	(0.014)	(0.015)
Populism × civil service reform in force	0.043^{***}	0.031^{**}	0.153^{**}
	(0.011)	(0.012)	(0.059)
Observations	3,861	3,861	3,861
States	50	50	50
Year FE	Yes	Yes	Yes
State FE	Yes	Yes	Yes
Controls	Yes	Yes	Yes
State trends	Yes	Yes	Yes
Within R^2	0.668	0.666	0.665

Notes: The dependent variable is ln(Per capita personal income). In column (1), Populism equals one when the governor's mean LLM score exceeds the 75th percentile of the 1866-2023 state-year distribution. In column (2), Populism equals one whenever the governor's mean LLM score is positive. In column (3), Populism is the continuous governor mean LLM score. Civil service reform is measured using `cs_ref_with_repeal`, which equals one when a civil service reform is in force after accounting for observed repeals through 2011. For post-2011 years, the 2011 reform status is carried forward. Controls include number of speeches, log token count, and party indicators. All specifications include year and state fixed effects and state-specific linear trends. Standard errors clustered at the state level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Here we define the populism dummy using the 75th percentile as opposed to the baseline 90th-percentile definition. The main result that independent bureaucracies attenuate the negative effect still holds.

Civil service reform repeal. Our baseline measure is repeal-adjusted civil-service reform, based on adoption dates from Vannoni et al. (2021) and Ash et al. (2022). Because repeal information is documented through 2011, the full-sample specification carries forward each state's 2011 reform status for later years. We therefore report two checks. First, we compare the baseline results with estimates using the absorbing civil-service reform indicator, which is coded as one from the year of adoption onward. Second, we restrict the sample to years before 2011, and then to 2000–2010, where repeal status is directly observed and speech coverage is near-complete. Table 9 reports results for the full sample and the pre-2011 subsample; both confirm that the baseline results are not driven by the absorbing coding assumption. We additionally report estimates for the 2000–2010 subsample, where panel coverage is near-complete (and we do not rely on the extrapolation of the repeal years), to address the concern that missingness in the earlier decades may be driving the results. Estimates are consistent across all three samples.

Table 9: Populism, civil service reform coding, and log per capita income

	No repeal	Repeal-adjusted civil service reform	
	Full sample	Pre-2011	2000–2010
	(1)	(2)	(3)
High Populism	-0.0577^{***}	-0.0380^{***}	-0.0291^{***}
	(0.0212)	(0.0135)	(0.0069)
Civil service reform	−0.0152	0.0040	0.0115
	(0.0172)	(0.0126)	(0.0105)
High Populism × Civil service reform	0.0627^{***}	0.0413^{***}	0.0201^{**}
	(0.0214)	(0.0144)	(0.0096)
Observations	3,861	3,213	547
States	50	50	50
Year FE	Yes	Yes	Yes
State FE	Yes	Yes	Yes
State-specific trends	Yes	Yes	Yes
Controls	Yes	Yes	Yes
Sample	Full	Year $<$ 2011	2000 $\leq$ Year $<$ 2011
Within R^2	0.666	0.680	0.744

Notes: The dependent variable is ln(Per capita personal income). High Populism is the indicator *Pop*, equal to one when the governor's mean LLM score exceeds the 90th percentile of the 1866-2023 state-year distribution. Column (1) uses the civil service reform indicator without accounting for repeals, coded as one from the year of adoption onward. Columns (2)–(3) use `cs_ref_with_repeal`, which equals one when a civil service reform is in force after accounting for observed repeals. Column (2) restricts the sample to years before 2011, so it does not rely on the post-2011 carry-forward assumption. Column (3) further restricts the sample to 2000–2010, where missingness in the repeal-adjusted coding is minimal (547 of 550 state-years) and no post-2011 carry-forward assumption is required. Controls include number of speeches, log token count, and party indicators. All specifications include year and state fixed effects and state-specific linear trends. Standard errors clustered at the state level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

B.1 Additional controls

Table 10: Populism, civil service reform, and log per capita income: additional controls and lagged dependent variable

	Civil service reform, accounting for repeals	
	Additional controls	Additional controls + lagged y
	(1)	(2)
High Populism	-0.034^{***}	-0.013^{***}
	(0.012)	(0.003)
Civil service reform in force	0.008	-0.004
	(0.014)	(0.003)
High Populism × civil service reform in force	0.047^{***}	0.015^{***}
	(0.013)	(0.004)
$L.\ln(y)$		0.816^{***}
		(0.030)
Observations	3,861	3,813
States	50	50
Year FE	Yes	Yes
State FE	Yes	Yes
Baseline controls	Yes	Yes
Additional controls	Yes	Yes
State trends	Yes	Yes
Lagged y	No	Yes
Within R^2	0.677	0.890

Notes: The dependent variable is ln(Per capita personal income). High Populism is the indicator *Pop*, equal to one when the governor's mean LLM score exceeds the 1866-2023 90th percentile of the state-year distribution. Civil service reform is measured using `cs_ref_with_repeal`, which equals one when a civil service reform is in force after accounting for observed repeals through 2011. For post-2011 years, the 2011 reform status is carried forward. Baseline controls include number of speeches, log token count, and party indicators. Additional controls are ln(percentBlack), ln(percentForeignBorn), ln(firms), and ln(bankruptcies), drawn from the Correlates of State Policy Project (Grossmann et al., 2021) and completed by interpolation and extrapolation within state before taking logs; see Appendix H.5 for details. Column (2) additionally includes the first lag of the dependent variable. All specifications include year and state fixed effects and state-specific linear trends. Standard errors clustered at the state level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

B.2 Lagged explanatory variables

Table 11: Populism, civil service reform, and per capita income: alternative specifications

	ln(per capita income)		Per capita income	Δ ln(per capita income)
	Current RHS	Lagged RHS	Current RHS	Current RHS
	(1)	(2)	(3)	(4)
High Populism	-0.0380^{***}	-0.0360^{***}	-626.53^{*}	-0.0083^{***}
	(0.0116)	(0.0105)	(319.67)	(0.0022)
Civil service reform in force	0.0047	-0.0010	-271.44	-0.0060^{***}
	(0.0147)	(0.0153)	(360.70)	(0.0017)
High Populism $\times$ Civil service reform in force	0.0490^{***}	0.0481^{***}	$1{,}091.13^{**}$	0.0084^{***}
	(0.0126)	(0.0117)	(417.03)	(0.0028)
Observations	3,861	3,811	3,861	3,813
States	50	50	50	50
Year FE	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes
State-specific trends	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
RHS variables lagged	No	Yes	No	No
Within R^2	0.666	0.661	0.733	0.009

Notes: Columns (1)–(2) use ln(Per capita personal income) as the dependent variable; column (3) uses per capita personal income in levels; column (4) uses the first difference of log per capita personal income, Δ ln(Per capita personal income). High Populism is the indicator *Pop*, equal to one when the governor's mean LLM score exceeds the 90th percentile of the state-year distribution. Civil service reform is measured using `cs_ref_with_repeal`, which equals one when a civil service reform is in force after accounting for observed repeals. Columns (1), (3), and (4) use contemporaneous right-hand-side variables. Column (2) lags High Populism, civil service reform, their interaction, number of speeches, log token count, and party indicators by one year, addressing reverse-causality concerns. Column (4) addresses concerns about non-stationarity in ln(Per capita personal income) by estimating in growth rates. All specifications include year and state fixed effects and state-specific linear trends. Standard errors clustered at the state level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C Main Difference-In-Difference results

Table 12: Strict-missing DiD estimates: high populism's effect on log per-capita income, by modern civil-service regime

	`B_modern` = 0 (no civil-service reform in force)			`B_modern` = 1 (civil-service reform in force)		
	Estimate	SE	N/S	Estimate	SE	N/S
Pre-treatment placebos						
Placebo_4	0.009	0.006	35/5	0.003	0.004	69/11
Placebo_3	0.007	0.005	38/5	0.002	0.003	74/11
Placebo_2	0.001	0.006	39/5	−0.002	0.003	79/11
Placebo_1	−0.003	0.006	42/5	−0.005	0.004	81/11
Joint nullity test (placebos):		$p = 0.214$			$p = 0.309$	
Post-treatment effects						
Effect_1	−0.004	0.007	57/8	0.002	0.004	111/13
Effect_2	-0.011^{**}	0.006	54/8	0.005	0.004	109/13
Effect_3	-0.008^{*}	0.005	53/8	0.003	0.003	104/13
Effect_4	-0.006^{*}	0.003	50/8	0.003	0.002	99/13
Joint nullity test (effects):		$p = 0.012$			$p = 0.588$	
Av_tot_eff	-0.018^{*}	0.010	148/32	0.008	0.007	320/52
Avg. exposure (years)		2.50			2.54	

Notes: Estimates from `did_multiplegt_dyn` (de Chaisemartin et al., 2025) on the panel of U.S. states from 2010 onwards. The treatment is a binary indicator for high populism, `p90`, equal to one when the governor's mean LLM populism score exceeds the 90th percentile of the state-year distribution of governor-mean LLM scores. The moderator `B_modern` separates states by their modern civil-service regime: `B_modern`=0 denotes states without civil-service reform in force during the modern period, while `B_modern`=1 denotes states with civil-service reform in force throughout the modern period. Standard errors are clustered at the state level. The specification uses the `normalized`, `same_switchers`, `same_switchers_pl`, and `drop_if_d_miss_before_first_switch` options. We refer to this as the strict-missing specification because groups with missing treatment status before their first switch are dropped. In the event-study and placebo rows, N/S reports the number of observations and the number of switchers contributing to each horizon. In the Av_tot_eff row, N/S reports the number of observations and the number of switcher-periods. Av_tot_eff is the average cumulative total effect per unit of treatment, expressed in log points; "Avg. exposure" reports the average number of post-treatment periods over which the cumulative effect is computed. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$; based on cluster-robust standard errors.

Table 13: Event-study estimates: high populism's effect on log per-capita income, by modern civil-service regime (switchers into high populism only)

	`B_modern` = 0 (no civil-service reform in force)			`B_modern` = 1 (civil-service reform in force)		
	Estimate	SE	N/S	Estimate	SE	N/S
Pre-treatment placebos						
Placebo_4	0.009	0.006	35/5	0.005	0.005	65/8
Placebo_3	0.007	0.005	38/5	0.005	0.004	69/8
Placebo_2	0.001	0.006	39/5	0.000	0.004	74/8
Placebo_1	−0.003	0.006	42/5	−0.003	0.004	75/8
Joint nullity test (placebos):		$p = 0.214$			$p = 0.036$	
Post-treatment effects						
Effect_1	−0.004	0.007	57/8	0.002	0.004	105/10
Effect_2	-0.011^{**}	0.006	54/8	0.003	0.004	104/10
Effect_3	-0.008^{*}	0.005	53/8	0.002	0.004	99/10
Effect_4	-0.006^{*}	0.003	50/8	0.002	0.003	95/10
Joint nullity test (effects):		$p = 0.012$			$p = 0.849$	
Av_tot_eff	-0.018^{*}	0.010	148/32	0.006	0.008	302/40
Avg. exposure (years)		2.50			2.55	

Notes: Estimates from `did_multiplegt_dyn` (de Chaisemartin et al., 2025) on the panel of U.S. states from 2010 onwards. The treatment is a binary indicator for high populism, equal to one when the governor's mean LLM populism score exceeds the 90th percentile of the state-year distribution of governor-mean LLM scores. The moderator `B_modern` separates states by their modern civil-service regime. Standard errors are clustered at the state level. This specification uses the `normalized`, `same_switchers`, and `same_switchers_pl` options, and additionally restricts the analysis to switches into treatment with `switchers(in)`. In the event-study and placebo rows, N/S reports the number of observations and the number of switchers contributing to each horizon. In the Av_tot_eff row, N/S reports the number of observations and the number of switcher-periods. Av_tot_eff is the average cumulative total effect per unit of treatment, expressed in log points; "Avg. exposure" reports the average number of post-treatment periods over which the cumulative effect is computed. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$; based on cluster-robust standard errors.

Table 14: Event-study estimates: high populism's effect on log per-capita income, by modern civil-service regime (all switchers)

	`B_modern` = 0 (no civil-service reform in force)			`B_modern` = 1 (civil-service reform in force)		
	Estimate	SE	N/S	Estimate	SE	N/S
Pre-treatment placebos						
Placebo_4	0.009	0.006	35/5	0.003	0.004	72/11
Placebo_3	0.007	0.005	38/5	0.003	0.003	77/11
Placebo_2	0.001	0.006	39/5	−0.001	0.003	82/11
Placebo_1	−0.003	0.006	42/5	−0.006	0.004	104/14
Joint nullity test (placebos):		$p = 0.214$			$p = 0.112$	
Post-treatment effects						
Effect_1	−0.004	0.007	57/8	0.001	0.003	134/16
Effect_2	-0.011^{**}	0.006	54/8	0.004	0.004	112/13
Effect_3	-0.008^{*}	0.005	53/8	0.003	0.003	107/13
Effect_4	-0.006^{*}	0.003	50/8	0.003	0.002	102/13
Joint nullity test (effects):		$p = 0.012$			$p = 0.548$	
Av_tot_eff	-0.018^{*}	0.010	148/32	0.008	0.007	330/55
Avg. exposure (years)		2.50			2.45	

Notes: Estimates from `did_multiplegt_dyn` (de Chaisemartin et al., 2025) on the panel of U.S. states from 2010 onwards. The treatment is a binary indicator for high populism, equal to one when the governor's mean LLM populism score exceeds the 90th percentile of the state-year distribution of governor-mean LLM scores. The moderator `B_modern` separates states by their modern civil-service regime. Standard errors are clustered at the state level. This specification uses the `normalized` option but does not impose the `same_switchers` or `same_switchers_pl` restrictions, so the set of switchers contributing to each event-study and placebo horizon may vary. In the event-study and placebo rows, N/S reports the number of observations and the number of switchers contributing to each horizon. In the Av_tot_eff row, N/S reports the number of observations and the number of switcher-periods. Av_tot_eff is the average cumulative total effect per unit of treatment, expressed in log points; "Avg. exposure" reports the average number of post-treatment periods over which the cumulative effect is computed. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$; based on cluster-robust standard errors.

Table 15: Event-study estimates: high populism's effect on per-capita income, by modern civil-service regime

	`B_modern` = 0 (no civil-service reform in force)			`B_modern` = 1 (civil-service reform in force)		
	Estimate	SE	N/S	Estimate	SE	N/S
Pre-treatment placebos						
Placebo_4	424.634	302.696	35/5	79.311	199.993	69/11
Placebo_3	338.839	231.127	38/5	68.776	178.149	74/11
Placebo_2	90.767	250.488	39/5	−152.398	199.507	79/11
Placebo_1	−94.492	306.049	42/5	−269.880	218.214	81/11
Joint nullity test (placebos):	$p = 0.176$			$p = 0.204$		
Post-treatment effects						
Effect_1	−325.693	319.081	57/8	61.536	171.559	111/13
Effect_2	−709.435**	278.167	54/8	259.817	205.400	109/13
Effect_3	−538.323**	241.944	53/8	300.381	195.143	104/13
Effect_4	−451.656**	191.845	50/8	255.467	169.415	99/13
Joint nullity test (effects):	$p = 0.005$			$p = 0.555$		
Av_tot_eff	−1,291.540**	555.077	148/32	634.861	417.211	320/52
Avg. exposure (years)	2.50			2.54		

Notes: Estimates from `did_multiplegt_dyn` (de Chaisemartin et al., 2025) on the panel of U.S. states from 2010 onwards. The outcome is per-capita personal income in levels. The treatment is a binary indicator for high populism, equal to one when the governor's mean LLM populism score exceeds the 90th percentile of the state-year distribution of governor-mean LLM scores. The moderator `B_modern` separates states by their modern civil-service regime. Standard errors are clustered at the state level. This specification uses the `normalized`, `same_switchers`, `same_switchers_pl`, and `drop_if_d_miss_before_first_switch` options. In the event-study and placebo rows, N/S reports the number of observations and the number of switchers contributing to each horizon. In the Av_tot_eff row, N/S reports the number of observations and the number of switcher-periods. Av_tot_eff is the average cumulative total effect per unit of treatment, expressed in income levels; "Avg. exposure" reports the average number of post-treatment periods over which the cumulative effect is computed. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$; based on cluster-robust standard errors.

D Populism Score Validation

D.1 Dictionary-based populism scores

We construct two dictionary-based measures of populist rhetoric in governors' State of the State speeches[18]: `populism_final` and `populism_pauwels`. Both measures are computed at the speech level, aggregated to the governor level, and then merged onto the state-year panel. `populism_final` uses the dictionary implementation in Gennaro et al. (2024), which builds on Pauwels (2011) by expanding the term list using WordNet, stemming the resulting terms, manually removing clearly irrelevant entries, and separating the remaining terms into anti-elite and people-centric components. `populism_pauwels`, by contrast, uses the original Pauwels dictionary, which we split into the same two conceptual components for comparability. In both cases, the measure is constructed to capture populist rhetoric only when both the anti-elite and people-centric dimensions are present.

Each speech is lowercased, tokenized using alphabetic tokens only, stripped of English stopwords, and Porter-stemmed. The dictionary scores are computed on a trimmed corpus that excludes speeches below the first percentile and above the ninety-ninth percentile of the processed token-count distribution.[19] In the corrected scoring pipeline, the full combined corpus contains 3,355 speeches; the trimmed corpus contains 3,287 speeches, dropping 68 extreme-length documents. TF–IDF document frequencies are then computed on this trimmed corpus.

Let d index speeches, let n_d be the number of processed tokens in speech d, and let c_{dj} be the count of dictionary stem j in speech d. For a given dictionary $k \in \{\text{final}, \text{pauwels}\}$, document frequency is

$$df_j = \sum_{d=1}^{N} \mathbf{1}\{c_{dj} > 0\},$$

where N is the number of speeches in the trimmed corpus. The TF–IDF weight for term j in speech d is

$$w_{dj} = \frac{c_{dj}}{n_d} \log\left(\frac{N}{df_j}\right).$$

Each dictionary is divided into an anti-elite component D_k^E and a people-centric component D_k^P. We compute component scores as

$$Elite_{dk} = \sum_{j \in D_k^E} w_{dj}, \qquad People_{dk} = \sum_{j \in D_k^P} w_{dj}.$$

Populism is then coded using a non-compensatory rule: a speech receives a positive populism score only when both the anti-elite and people-centric components are present. Thus,

$$Populism_{dk} = \begin{cases} Elite_{dk} + People_{dk}, & \text{if } Elite_{dk} > 0 \text{ and } People_{dk} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

We construct two populism scores, `populism_final` and `populism_pauwels`, each based on a separate dictionary. The stemmed tokens for both are listed in Tables 16 and 17, respectively, with terms divided into anti-elite (D^E) and people-centric (D^P) stems.

[18] We use "State of the State speeches" as shorthand for the governors' State of the State corpus. The corpus is centered on annual gubernatorial addresses, but some source files include closely related budget addresses, budget-session remarks, or legislative-session materials.

[19] We trim these tails because the shortest files are often incomplete or otherwise too limited to provide a reliable measure of rhetorical style, while the longest files frequently include material beyond the governor's address itself, such as extended records of the surrounding legislative day's proceedings.

Table 16: Stemmed tokens in the cleaned dictionary (`populism_final`)

Anti-elite stems (D^E)	**People-centric stems (D^P)**
absurd, absurdli, admit, admitt	peopl
arrog, arrogantli, betrai	tradit, tradition
cast, class, corrupt, deceit	direct, directli
directori, elit, elitist	referendum
establish, polit, politic	
politician, promin, promis	
propaganda, regim, regimen	
rule, shame, treason	
undemocrat	

Table 17: Stemmed tokens in the Pauwels dictionary (`populism_pauwels`)

Anti-elite stems (D^E_{pauwels})	**People-centric stems (D^P_{pauwels})**
absurd, admit, arrogant, betray	people
capitul, caste, class, corrupt	tradition
deceit, establishm, mafia	direct
particrat, politic, promis	referend
promise, propaganda, regime	
ruling, shame, shameless, treason	
undemocratic	

D.1.1 Validation diagnostics

The diagnostics below use the corrected, trimmed speech-level corpus. With the cleaned dictionary, 99.9 percent of speeches contain at least one dictionary token in either the anti-elite or people-centric dimension, and 98.4 percent contain tokens from both dimensions. Under the non-compensatory rule, this means that 98.4 percent of speeches receive a positive `populism_final` score. The corresponding shares for the Pauwels dictionary are 99.9 percent for either dimension and 94.8 percent for both dimensions.

Dictionary hits are not sparse in the trimmed corpus: only 0.15 percent of speeches contain at most two cleaned-dictionary populist tokens, and only 1.37 percent contain at most five. The mean cleaned-dictionary token share is 0.0100, compared with 0.0087 for the Pauwels dictionary. Because binary coverage is near-universal, our interpretation relies on the continuous TF–IDF intensity of the anti-elite and people-centric components, not merely on whether a speech contains any dictionary word.

The `populism_final` score is moderately correlated with the raw share of cleaned-dictionary tokens ($r = 0.515$) and effectively uncorrelated with processed speech length ($r = 0.001$). This suggests that the TF–IDF measure captures variation in the relative intensity of populist language, rather than mechanical differences in document length. We treat `populism_final` and `populism_pauwels` as complementary dictionary-based measures. The former is constructed from the cleaned and conceptually split dictionary, while the latter serves as a robustness check based on the original Pauwels vocabulary.

D.1.2 Aggregation to the state-year panel

The dictionary measures are first computed for each speech. We then aggregate speech-level scores to the governor level by taking the mean across available speeches for each governor

within a state. The resulting governor-level scores are assigned to all state-years in which that governor is in office. This aggregation means that trimming a particular speech does not necessarily remove that governor from the state-year panel if the same governor has another speech retained in the trimmed corpus.

In the corrected data, 68 speeches are dropped by trimming speeches below the 1st percentile and above the 99th percentile of token counts. Sixty-four of these have another retained speech for the same normalized state-governor pair. The four exceptions are William S. Taylor in Kentucky, Walter Maddock in North Dakota, Coe I. Crawford in South Dakota, and Oscar Branch Colquitt in Texas. These exceptions do not change the estimation sample in the main regressions: the relevant South Dakota and Texas state-years have missing `populism_final` but also missing income data, and therefore do not enter the regression sample.

D.1.3 Dictionary based results

Table 18 examines whether the baseline results are robust to replacing the LLM-based populism measure with dictionary-based alternatives described in Appendix D.1. These measures differ fundamentally from the LLM-based score in their construction: rather than leveraging a language model's contextual understanding of political rhetoric, they count the frequency of pre-specified stems associated with anti-elite and people-centric discourse. As such, they are considerably noisier proxies for populist rhetoric, which motivates the use of the LLM-based measure in the main analysis. Panel A uses the continuous dictionary scores directly. Despite the change in measurement approach, the interaction with bureaucratic capacity remains positive across all four specifications and is statistically significant in the civil service reform columns, mirroring the pattern in the baseline. The main effect of the populism score is consistently negative, in line with the main results. The Pauwels dictionary yields somewhat larger and more precisely estimated coefficients throughout, which may reflect its broader vocabulary capturing a wider range of populist rhetoric. Panels B and C binarise the dictionary scores at the 75th and 90th percentiles of the state-year distribution respectively, matching the indicator construction used in the main results. The interaction terms are again consistently positive and grow larger as the threshold tightens. At the 90th percentile, all interaction terms reach conventional levels of statistical significance across both dictionaries and both bureaucratic capacity measures, providing the clearest dictionary-based replication of the baseline finding. Overall, the finding that populism is associated with lower income in low-capacity states but not in high-capacity states appears robust to the choice of measurement approach, suggesting it reflects a substantive feature of the data rather than an artefact of the LLM scoring methodology. That said, the greater noise inherent in dictionary-based approaches — reflected in wider standard errors throughout — underscores the value of the LLM measure for the main analysis.

D.2 GPT-5.5-based populism scores

We re-score the full SOTS corpus with GPT-5.5 using the identical holistic-grading prompt and pipeline described above for Qwen, and construct an analogous binary treatment indicator at the 90th percentile of the 1866-2023 state-year score distribution (cutoff 0.25). Validation against the human-coded sample is in Appendix 25: Qwen tracks the human ratings more closely, while GPT-5.5 exhibits a substantial upward bias, so Qwen remains our primary measure. Furthermore, for 34 speeches the model did not return a parseable holistic-reasoning object on the standard call; these were recovered with a score-only fallback that returns the scalar populism grade without the accompanying reasoning text. The score-only fallback skips the holistic-reasoning step used for the other speeches.

We first replace the Qwen indicator with the GPT-5.5 indicator in the main fixed-effects interaction specification of Equation 5.0.1. Table 19 reports the estimates. The pattern is unchanged and, if anything, slightly sharper than in the Qwen baseline: in the full specification

Table 18: Dictionary-based populism, civil service reform, and log per capita income

	Civil service reform, accounting for repeals	
	Gennaro et al. (2024) dict.	Pauwels (2011) dict.
	(1)	(2)
Panel A: Continuous dictionary-based populism score		
Populism score	−2.156	−6.393*
	(2.114)	(3.198)
Civil service reform in force	−0.010	−0.011
	(0.018)	(0.018)
Populism score × civil service reform in force	5.928**	8.763**
	(2.740)	(3.586)
Panel B: High dictionary-based populism, 75th percentile threshold		
High populism	−0.006	−0.026**
	(0.009)	(0.011)
Civil service reform in force	0.006	0.003
	(0.015)	(0.015)
High populism × civil service reform in force	0.014	0.027**
	(0.010)	(0.011)
Panel C: High dictionary-based populism, 90th percentile threshold		
High populism	−0.011	−0.020*
	(0.012)	(0.012)
Civil service reform in force	0.007	0.006
	(0.015)	(0.015)
High populism × civil service reform in force	0.034**	0.035***
	(0.013)	(0.013)
Observations	3,861	3,861
States	50	50
Year FE	Yes	Yes
State FE	Yes	Yes
Controls	Yes	Yes
State trends	Yes	Yes

Notes: The dependent variable is ln(Per capita personal income). Panel A uses continuous dictionary-based populism scores. Panels B and C use binary high-populism indicators equal to one when the governor's dictionary-based populism score exceeds the 75th and 90th percentiles, respectively, of the 1866-2023 state-year distribution. Column (1) uses the cleaned dictionary based on Gennaro et al. (2024); column (2) uses the dictionary from Pauwels (2011). Civil service reform is measured using `cs_ref_with_repeal`, which equals one when a civil service reform is in force after accounting for observed repeals through 2011. For post-2011 years, the 2011 reform status is carried forward. All specifications include number of speeches, log token count, party indicators, year and state fixed effects, and state-specific linear trends. Standard errors clustered at the state level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

(column 3) high-populism governors are associated with roughly 5.6% lower per-capita income where civil-service reform is not in force ($\hat{\beta}_1 = -0.056$, $p < 0.01$), the interaction with reform in force is positive and larger in magnitude ($\hat{\beta}_3 = 0.062$, $p < 0.01$). The sign pattern — negative main effect, offsetting positive interaction — holds across all three columns.

Table 19: Populism (GPT-5.5), civil service reform, and log per capita income

	Civil service reform, accounting for repeals		
	(1)	(2)	(3)
High populism	-0.132^{***}	-0.057^{***}	-0.056^{***}
	(0.039)	(0.013)	(0.012)
Civil service reform in force	-0.009	0.006	0.005
	(0.023)	(0.014)	(0.014)
High populism × civil service reform in force	0.118^{**}	0.059^{***}	0.062^{***}
	(0.048)	(0.017)	(0.015)
Observations	3,861	3,861	3,861
States	50	50	50
Year FE	Yes	Yes	Yes
State FE	Yes	Yes	Yes
Controls	No	No	Yes
State trends	No	Yes	Yes
Within R^2	0.043	0.665	0.669

Notes: The dependent variable is ln(Per capita personal income). High populism is a binary indicator equal to one when the governor's mean GPT-5.5 populism score exceeds the 90th percentile of the 1866-2023 state-year distribution (cutoff 0.25). Civil service reform is measured using `cs_ref_with_repeal`, which equals one when a civil service reform is in force after accounting for observed repeals through 2011. For post-2011 years, the 2011 reform status is carried forward. Controls include number of speeches, log token count, and party indicators. Standard errors clustered at the state level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

We then re-estimate the heterogeneity-robust dynamic DiD of Section 7 using the GPT-5.5 indicator. Table 20 and Figure 7 report the normalized event-study estimates by modern civil-service regime. The directional pattern matches the Qwen-based results: in states without civil-service reform in force (`B_modern` $= 0$) the post-treatment coefficients are negative at every horizon and jointly significant ($p = 0.035$), with an average total effect of -0.010 log points; in states with reform in force (`B_modern` $= 1$) the effects are essentially zero (joint $p = 0.99$; average total effect 0.000).

Two caveats temper this evidence. First, the magnitudes are smaller than in the Qwen estimates, and neither the individual post-treatment coefficients nor the average total effect is individually significant. Second, the pre-treatment placebos require care: the placebo F-test rejects joint nullity in the no-reform group ($p < 0.001$). The only individually significant coefficient in the pre-treatment period is the second lag, which, which is marginally significant at the 10% significance level. We therefore read the GPT-based DiD as directionally consistent with our main results, while noting that the smaller, individually-insignificant coefficients make it weaker corroboration than the Qwen estimates.

Table 20: GPT-based event-study estimates: high populism's effect on log per-capita income, by modern civil-service regime

	`B_modern` $= 0$ (no civil-service reform in force)			`B_modern` $= 1$ (civil-service reform in force)		
	Estimate	SE	N/S	Estimate	SE	N/S
Pre-treatment placebos						
Placebo_4	0.002	0.015	27/3	0.000	0.003	104/5
Placebo_3	−0.000	0.011	27/3	0.000	0.004	106/5
Placebo_2	−0.010*	0.005	27/3	−0.002	0.004	108/5
Placebo_1	−0.006	0.005	27/3	−0.004	0.004	109/5
Joint nullity test (placebos):		$p < 0.001$			$p = 0.419$	
Post-treatment effects						
Effect_1	−0.004	0.007	42/4	0.000	0.004	143/8
Effect_2	−0.011	0.007	42/4	0.000	0.005	142/8
Effect_3	−0.004	0.004	42/4	0.000	0.004	140/8
Effect_4	−0.001	0.002	42/4	−0.000	0.004	138/8
Joint nullity test (effects):		$p = 0.035$			$p = 0.991$	
Av_tot_eff	−0.010	0.009	168/16	0.000	0.010	390/32
Avg. exposure (years)		2.50			2.57	

Notes: Estimates from `did_multiplegt_dyn` (de Chaisemartin et al., 2025) on the panel of U.S. states from 2010 onwards. The treatment is a binary indicator for high populism constructed from GPT scores, equal to one when the governor's mean GPT score exceeds the 90th percentile of the 1866–2023 state-year distribution of governor-mean GPT scores. The cutoff is 0.25. The moderator `B_modern` separates states by their modern civil-service regime. Standard errors are clustered at the state level. The specification uses the `normalized`, `same_switchers`, and `same_switchers_pl` options to keep the composition of switchers fixed across event-study and placebo horizons. In the event-study and placebo rows, N/S reports the number of observations and the number of switchers contributing to each horizon. The number of switchers can differ between placebo and post-treatment rows because some states have a full four years of post-treatment data within the 2010–2023 window but fewer than four pre-treatment years. The `same_switchers` and `same_switchers_pl` options hold composition fixed within the post-treatment and placebo windows separately. In the Av_tot_eff row, N/S reports the number of observations and the number of switcher-periods. Av_tot_eff is the average cumulative total effect per unit of treatment, expressed in log points; "Avg. exposure" reports the average number of post-treatment periods over which the cumulative effect is computed. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$; based on cluster-robust standard errors.

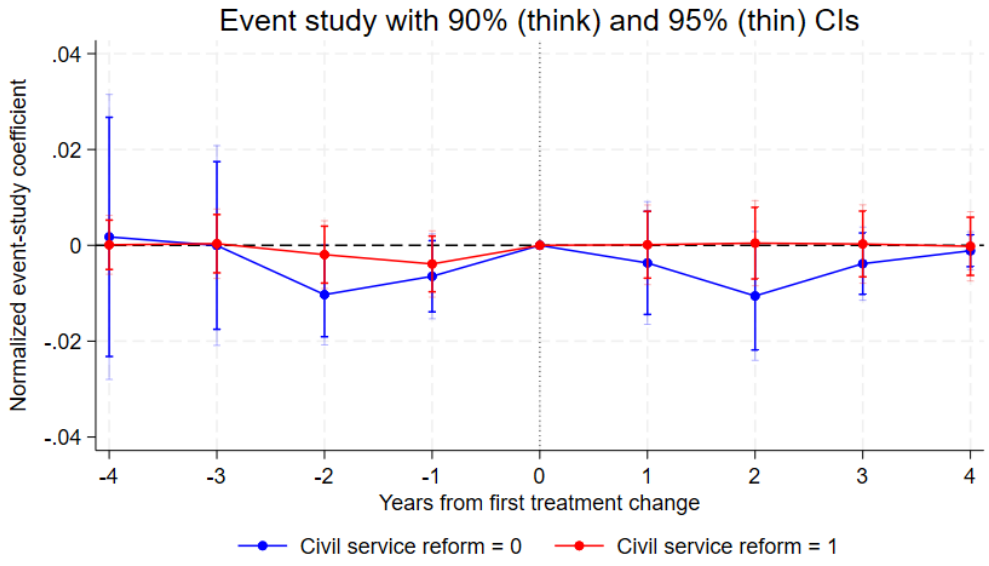


Figure 7: GPT-based event-study estimates of populism's effect on log per capita income, by pre-takeover civil-service reform status. Solid lines plot the normalized event-study coefficients $\widehat{DID}_\ell^{\,n}$ for $\ell = -4, \ldots, 4$, separately for states without civil-service reform in force during the modern era (`B_modern` $= 0$) and states with reform in force throughout (`B_modern` $= 1$). Thick (faint) vertical bars show cluster-robust 90% (95%) confidence intervals. The post-takeover trajectory is negative in the no-reform group and flat in the reform group, mirroring the Qwen-based estimates in Figure 5.

E Oster (2019) Bounds for Coefficient Stability

Framework

Following Oster (2019), we assess the robustness of our interaction-term estimates to omitted variable bias. The method compares a short regression with year and state fixed effects only to a full regression that adds controls and state-specific linear time trends. The movement in both the coefficient and the within R^2 is then used to infer the degree of selection on unobservables that would be required to explain the result away.

Notation. Let:

- $\tilde{\beta}$ and $\tilde{R}^2_{\text{within}}$: coefficient and within R^2 from the short regression with year and state fixed effects only;
- $\hat{\beta}$ and $\hat{R}^2_{\text{within}}$: coefficient and within R^2 from the full regression with controls and state-specific linear trends;
- $R_{\max}$: a hypothetical maximum R^2 attainable if all relevant unobservables were included.

Key formulae. The degree of proportional selection δ required to drive the true effect to zero is:

$$\delta^* = \frac{\hat{\beta}\,(\hat{R}^2_W - \tilde{R}^2_W)}{(\tilde{\beta} - \hat{\beta})\,(R_{\max} - \hat{R}^2_W)} \tag{E.0.1}$$

If $|\delta^*| > 1$, unobservables would need to be more important than observables to fully explain the coefficient, which is conventionally regarded as evidence of robustness.

The bias-adjusted coefficient under equal selection, $\delta = 1$, is:

$$\beta^*(\delta = 1) \;=\; \hat{\beta} \;-\; \frac{(\tilde{\beta} - \hat{\beta})\,(R_{\max} - \hat{R}^2_W)}{\hat{R}^2_W - \tilde{R}^2_W} \tag{E.0.2}$$

Choice of $R_{\max}$. Following Oster's (2019) recommendation, we set $R_{\max} = \min\{1,\ 1.3 \times \hat{R}^2_W\}$. All R^2 values are within R^2 from the fixed-effects regressions, which is the appropriate measure for models absorbing year and state fixed effects.

Oster Bounds for 90th-Percentile Populism Threshold (*Pop*)

Calculations for $p90$ interaction

Interaction	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{R}^2_W$	$\hat{R}^2_W$	$R_{\max}$
$p90 \times$ CS Reform in Force	0.0788	0.0490	0.0156	0.6664	0.8663

Worked calculation:

Pop × **CS Reform in Force:**

$$\begin{aligned}
\delta^* &= \frac{0.0490 \times (0.6664 - 0.0156)}{(0.0788 - 0.0490) \times (0.8663 - 0.6664)} \\
&= \frac{0.0319}{0.0059} \\
&= 5.35
\end{aligned}$$

$$\begin{aligned}
\beta^*(\delta{=}1) &= 0.0490 - \frac{(0.0788 - 0.0490) \times (0.8663 - 0.6664)}{0.6664 - 0.0156} \\
&= 0.0398
\end{aligned}$$

Summary

Oster Bounds: Summary of Results

Interaction	δ^*	$\beta^*(\delta{=}1)$	$\lvert\delta^*\rvert > 1$?
Pop × CS Reform in Force	5.35	0.0398	✓

The *Pop* × CS Reform in Force interaction satisfies $|\delta^*| > 1$. The bias-adjusted coefficient under equal selection, $\beta^*(\delta = 1)$, remains positive at 0.0398, compared with the full-model estimate of 0.0490. This suggests that the positive interaction is not easily explained away by selection on unobservables of the same magnitude as selection on the included observables and state-specific trends.

F Synthetic control results

F.1 Main Institutional Splitters

Our main splitter is a composite bureaucracy measure constructed from three V-Dem indicators.

F.2 Bureaucracy Measure

We construct an exploratory bureaucracy measure from three V-Dem country-year variables. This is not an official V-Dem index.[20] Rather, it is a transparent composite based on three released V-Dem indicators that capture core features of the functioning of the public administration.

The three inputs are:

1. `v2clrspct`, *Rigorous and impartial public administration.* In the V-Dem codebook, this variable is defined by the question: *"Are public officials rigorous and impartial in the performance of their duties?"*

2. `v2stcritrecadm`, *Criteria for appointment decisions in the state administration.* V-Dem defines it by the question: *"To what extent are appointment decisions in the state administration based on personal and political connections, as opposed to skills and merit?"*

3. `v2strenadm`, *Bureaucratic remuneration.* V-Dem defines it by the question: *"To what extent are state administrators salaried employees?"*

Table 21: Bureaucracy Index and Component Correlations

	[1] `bureaucracy`	[2] `v2clrspct`	[3] `v2stcritrecadm`	[4] `v2strenadm`
`bureaucracy`	1			
`v2clrspct`	0.88	1		
`v2stcritrecadm`	0.88	0.71	1	
`v2strenadm`	0.76	0.49	0.48	1

Notes: `bureaucracy` is the first principal component of the three V-Dem components. `v2clrspct` is rigorous and impartial public administration; `v2stcritrecadm` is merit-based state recruitment; `v2strenadm` is bureaucratic remuneration. The index is computed on complete country-year observations in the GDP/V-Dem analysis panel and oriented so that larger values indicate stronger bureaucracy. This table is not restricted to the 28 treated events; it uses the full complete-case country-year panel. The treated-event sample and missing event-lag values are reported separately in the core-event-lag table. Correlations are Pearson correlations. Minimum pairwise $N = 7535$. All off-diagonal correlations are statistically significant at the 99% confidence level.

For each country-year (c, t), let

$$x_{1ct} = \texttt{v2clrspct}_{ct}, \quad x_{2ct} = \texttt{v2stcritrecadm}_{ct}, \quad x_{3ct} = \texttt{v2strenadm}_{ct}.$$

We first standardize each component across complete country-year observations:

$$z_{kct} = \frac{x_{kct} - \mu_k}{s_k},$$

where μ_k and s_k denote the sample mean and standard deviation of component k.

Stacking the standardized observations into a matrix Z, we then compute a one-factor score using the first singular vector:

$$Z = U\Sigma V^\top, \qquad b_{ct} = z_{ct}^\top v_1,$$

[20] Unlike the official V-Dem indices, the bureaucracy measure is a custom composite constructed from already-aggregated country-year indicators. We do not observe the underlying coder-level scores or the full measurement structure used by V-Dem to estimate its Bayesian latent-variable indices. We therefore summarize the standardized inputs with a simple one-factor score rather than attempting to replicate V-Dem's full Bayesian latent-variable procedure.

Table 22: Bureaucracy Index and Component Correlations: Core Event Lags

	[1] bureaucracy	[2] v2clrspct	[3] v2stcritrecadm	[4] v2strenadm
bureaucracy	1			
v2clrspct	0.88	1		
v2stcritrecadm	0.91	0.68	1	
v2strenadm	0.83	0.58	0.67	1

Notes: bureaucracy is the first principal component of the three V-Dem components. v2clrspct is rigorous and impartial public administration; v2stcritrecadm is merit-based state recruitment; v2strenadm is bureaucratic remuneration. The Funke core GDP sample contains 28 treated events. This table uses each event's $t-1$ bureaucracy values, but four events have incomplete bureaucracy data at $t-1$: Peru 1985, Peru 1990, Slovakia 1990, and Japan 2001. The correlations are therefore computed on the 24 complete event-lag observations, which are the events used in the high/low bureaucracy split. Correlations are Pearson correlations. Minimum pairwise $N = 24$. All off-diagonal correlations are statistically significant at the 99% confidence level.

where v_1 is the first column of V. This is equivalent to a PC1-style one-factor summary of the three standardized components. We orient the sign so that higher values correspond to stronger bureaucracy:

$$\operatorname{corr}\left(b_{ct}, \frac{1}{3}\sum_{k=1}^{3} z_{kct}\right) > 0.$$

In the split-sample GDP analysis, we use the lagged event-level value of this composite, measured in the year before the populist takeover. Let T_i denote the takeover year for event i. We therefore define the event-level bureaucracy measure as

$$B_i = b_{c_i, T_i - 1}.$$

An event is classified as "high" bureaucracy if its lagged composite value lies above the median among usable treated events in the estimation sample, and "low" otherwise. In practice, four of the 28 core events lack a valid composite value at $t-1$ (Slovakia 1990; Peru 1985, 1990; Japan 2001) and are excluded from the split, yielding $N = 24$. Thus, the split compares populist episodes that begin under relatively stronger versus weaker pre-treatment bureaucracy.

F.3 Core GDP events in the bureaucracy analysis

Table 23: Core GDP Events Included in the Bureaucracy Analysis

No.	Country	Event year	Period	Bureaucracy	Main doppelgangers
1	Argentina	1946	1931-1961	low	CHE,NLD,POL
2	Argentina	1973	1958-1988	low	PER,THA,VEN
3	Argentina	1989	1974-2004	high	MEX,ROU,NZL
4	Argentina	2003	1988-2018	high	CHE,MEX,CAN
5	Bolivia	1952	1937-1967	low	PHL,PRT,THA
6	Brazil	1951	1936-1966	low	TWN,ISL,THA
7	Brazil	1990	1975-2005	low	TUR,POL,PRY
8	Chile	1952	1937-1967	low	TWN,SWE,PHL
9	Ecuador	1952	1937-1967	low	THA,CHE,POL
10	Ecuador	1960	1945-1975	low	IDN,PRT,PHL
11	Ecuador	1968	1953-1983	low	AUT,BRA,FIN
12	Ecuador	1996	1981-2011	low	CZE,JPN,ARG
13	India	1966	1951-1981	high	EGY,GBR,MYS
14	Israel	1996	1981-2011	high	HUN,FRA,MEX
15	Italy	1994	1979-2009	high	BOL,DNK,FRA
16	Italy	2001	1986-2016	high	BOL,IND,MEX
17	Japan	2001	1986-2016	missing	MYS,NOR,FIN
18	Mexico	1970	1955-1985	low	TUR,NLD,HUN
19	New Zealand	1975	1960-1990	high	CHE,THA,ECU
20	Peru	1985	1970-2000	missing	TUR,ARG,POL
21	Peru	1990	1975-2005	missing	ARG
22	Philippines	1998	1983-2013	high	POL,ROU,TUR
23	Slovakia	1990	1975-2005	missing	PHL,THA,NLD
24	South Korea	2003	1988-2018	high	ZAF,FRA,ISR
25	Taiwan	2000	1985-2015	high	RUS,AUS,NOR
26	Thailand	2001	1986-2016	high	FIN,ROU
27	Turkey	2003	1988-2018	high	MYS,CHE,ISR
28	Venezuela	1999	1984-2014	low	CAN,PER,FIN

Notes: The GDP synthetic-control path is successfully estimated for all 28 events. 'Bureaucracy split' reports the event's assignment in the high/low bureaucracy analysis using the $t-1$ bureaucracy score; 'missing' indicates that the event is estimated in the GDP SCM but excluded from the bureaucracy split because the $t-1$ bureaucracy components are incomplete. The missing bureaucracy cases are Peru 1985, Peru 1990, Slovakia 1990, and Japan 2001. 'Main doppelganger countries' lists up to the three positive-weight donor countries with the largest synthetic-control weights.

Table 24: Bureaucracy Component Split Assignments for Core GDP Events

No.	Country	Event year	Bureaucracy	Impartial admin.	Merit recruitment	Remuneration
1	Argentina	1946	low	low	low	low
2	Argentina	1973	low	low	low	low
3	Argentina	1989	high	high	low	low
4	Argentina	2003	high	high	low	high
5	Bolivia	1952	low	low	low	low
6	Brazil	1951	low	low	low	low
7	Brazil	1990	low	low	low	low
8	Chile	1952	low	high	low	low
9	Ecuador	1952	low	low	low	low
10	Ecuador	1960	low	low	low	low
11	Ecuador	1968	low	low	low	low
12	Ecuador	1996	low	low	low	low
13	India	1966	high	high	high	high
14	Israel	1996	high	high	high	high
15	Italy	1994	high	high	high	high
16	Italy	2001	high	high	high	high
17	Japan	2001	missing	high	missing	missing
18	Mexico	1970	low	low	low	high
19	New Zealand	1975	high	high	high	high
20	Peru	1985	missing	high	missing	missing
21	Peru	1990	missing	low	missing	missing
22	Philippines	1998	high	low	high	high
23	Slovakia	1990	missing	missing	missing	missing
24	South Korea	2003	high	high	high	high
25	Taiwan	2000	high	high	high	low
26	Thailand	2001	high	low	high	high
27	Turkey	2003	high	high	high	high
28	Venezuela	1999	low	low	high	high

Notes: Entries report each core GDP event's high/low assignment using the $t-1$ value of the bureaucracy composite and its three V-Dem components. High is assigned when the event's $t-1$ value is strictly above the treated-event median for that variable; low is assigned when it is at or below the median. 'Missing' means the relevant $t-1$ value is unavailable. `v2clrspct` is rigorous and impartial public administration; `v2stcritrecadm` is merit-based state recruitment; `v2strenadm` is bureaucratic remuneration. Median cutoffs are: Bureaucracy = 0.219; Impartial admin. = 0.524; Merit recruitment = 0.906; Remuneration = 0.801.

F.4 Time placebo

Following Funke et al. (2023), we conduct a time-placebo exercise in which we shift the start year of each populist episode five years earlier and re-estimate the synthetic control using this fictitious treatment date, here we set it to 5 years before the actual populist takeover. If the baseline treatment timing has a causal interpretation, we should not observe systematic divergence between treated and synthetic outcomes in the period prior to the actual takeover date. In all placebo plots, we continue to define the "above" and "below" groups using institutions measured at the real pre-treatment year $t-1$ (i.e., relative to the actual takeover), rather than re-defining groups based on the placebo start date.

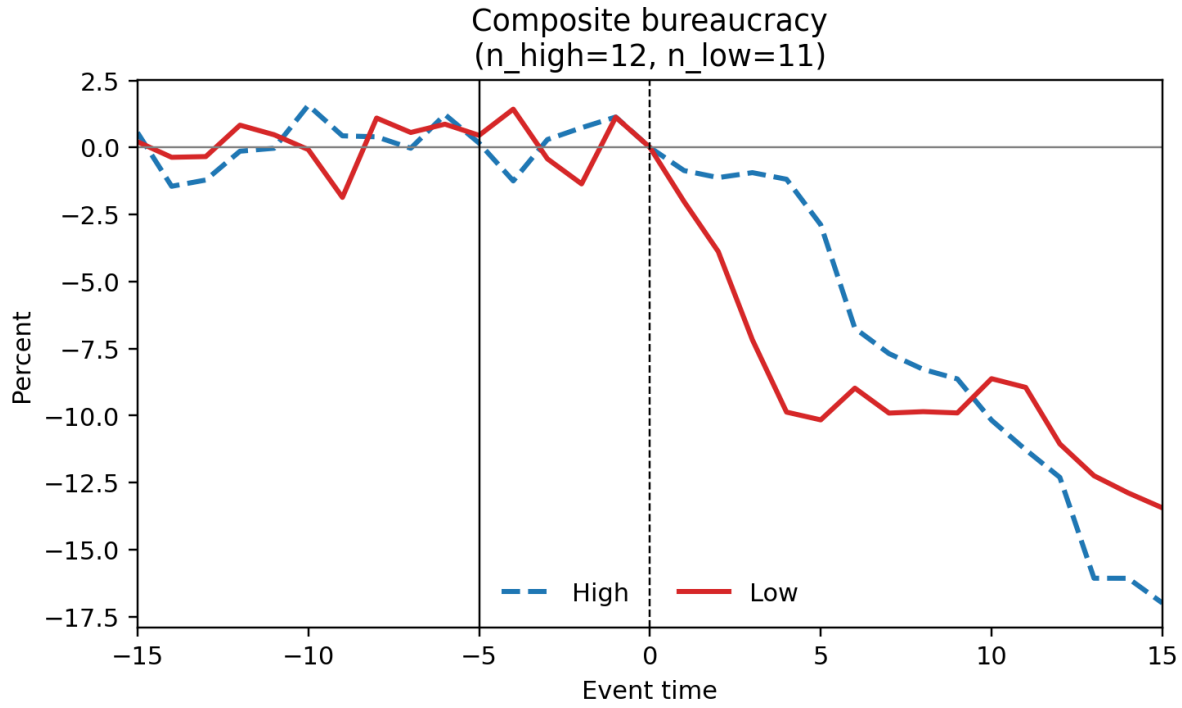


Figure 8: Time placebo: GDP-per-capita trajectories for treated episodes and their synthetic controls, split by pre-treatment composite bureaucracy (high vs. low), with the treatment date shifted five years earlier.

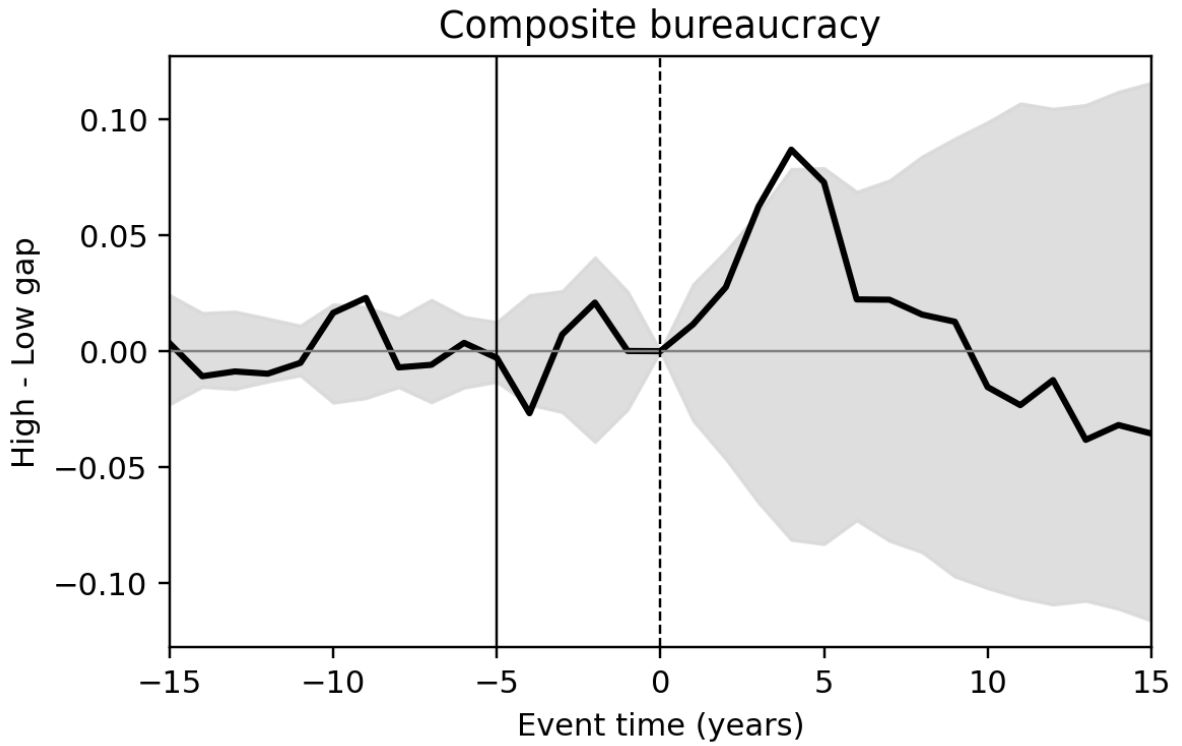


Figure 9: Time placebo: heterogeneity in post-"takeover" GDP-per-capita gaps by pre-treatment composite bureaucracy measure, with the treatment date shifted five years earlier. The high/low split is defined using composite bureaucracy measured at the real pre-treatment year $t-1$. The shaded region shows the pointwise 5th–95th percentile permutation envelope under the null of no heterogeneity.

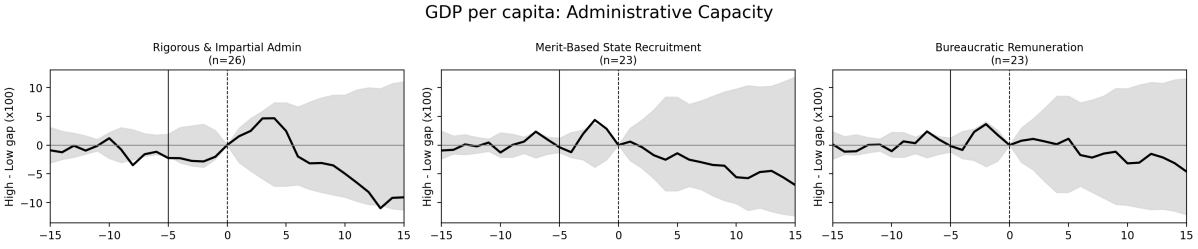


Figure 10: Time placebo: bureaucratic capacity (and components), with the treatment date shifted five years earlier. The high/low split uses institutions at the real pre treatment year. The shaded region shows the pointwise 5th–95th percentile permutation envelope under the null of no heterogeneity.

Across both the baseline figures and the time-placebo exercises, the observed high–low difference path remains inside the permutation envelope in nearly all cases throughout the pre-treatment period and during the placebo window between the fictitious and actual takeover dates. The main exception is the merit-based state-recruitment component in the five-year shift, for which the observed path briefly exits the envelope at some horizons; however, the corresponding split is not significant in the baseline.

A small number of placebo cases cannot be estimated in the strict implementation. In the $t = -5$ placebo, two of the 28 core events fail: Argentina 1946 and Slovakia 1990. Argentina 1946 is an extreme early case with war-era support problems, while Slovakia 1990 has unusually thin usable support. Once the placebo treatment boundary is moved back five years, these cases hit edge conditions in the strict `scpi` workflow and drop out. Importantly, this does not appear to be idiosyncratic to our implementation. In Funke et al. (2023)'s own Figure 8 placebo block, the hard-coded placebo case list contains 26 cases rather than the full 28, and the omitted cases line up with these same problematic episodes. Hence the realized placebo sample in our $t = -5$ exercise mirrors Funke et al.'s apparent effective sample.

F.5 Permutation-Based Randomization Test (Main-Text Inference)

We use a permutation-based randomization procedure for inference in the split-sample heterogeneity analysis. The permutation test asks whether the observed difference between the high- and low-institution groups is unusually large relative to what would arise under random reassignment of institutional labels.

The procedure is directly inspired by the placebo-based inference framework of Abadie et al. (2010), who assess significance of the treatment effect by reassigning the treatment label across units and comparing the treated unit's synthetic-control gap to the resulting placebo distribution. The treatment effect of populism on GDP has been established by Funke et al. (2023), and we apply the same logic one level up: rather than permuting the treated/control label, we permute the high/low institutional classification across treated episodes and assess whether the observed heterogeneity is extreme relative to the resulting null distribution.

Let $i = 1, \ldots, N$ index treated episodes and let $G_{i\tau}$ denote the episode-level synthetic-control gap at event time τ. These gap paths are taken as fixed inputs from the main estimation stage; no synthetic controls are re-estimated. Suppose the baseline split classifies N_H episodes as high-institution and $N_L = N - N_H$ as low-institution. The observed heterogeneity effect is

$$\Delta_\tau = \bar{G}_\tau^H - \bar{G}_\tau^L,$$

as defined in the preceding section.

Under the sharp null hypothesis that the institutional classification is uninformative—that is, that the treatment-effect paths $G_{i\tau}$ are exchangeable across the high and low groups—any assignment of N_H episodes to the high group and N_L to the low group is equally likely. We generate a permutation distribution by reshuffling the high/low labels $P = 1{,}000$ times while preserving the original group sizes. In each replication p, we compute

$$\Delta_\tau^{\pi(p)} = \bar{G}_\tau^{H^{\pi(p)}} - \bar{G}_\tau^{L^{\pi(p)}},$$

where $H^{\pi(p)}$ and $L^{\pi(p)}$ denote the pseudo-groups induced by permutation p. A fixed random seed ensures reproducibility.

We summarise the results by plotting the observed Δ_τ path against the 5th and 95th percentile envelope of the permutation distribution.

$$\left[q_{0.05,\tau},\ q_{0.95,\tau}\right].$$

Where the observed path falls outside this envelope, the null of no heterogeneity is rejected at the 10% significance level.[21]

[21] Note that, unlike the original placebo inference in Abadie et al. (2010), we do not normalise the test statistic

F.6 Bootstrap Confidence Bands (Alternative Inference)

We also compute episode-level bootstrap confidence bands for the split-sample heterogeneity paths. The unit of resampling is the treated episode (not the country-year observation). For each bootstrap replication $b = 1, \ldots, B$, we resample episodes with replacement within the high and low groups, recompute the group-average gaps, and then recompute the high minus-low difference path $\Delta_\tau^{*(b)}$. We set $B = 1000$ and construct pointwise 90% confidence bands using percentile bootstrap quantiles.

The bootstrap bands serve a different purpose from the permutation envelopes in the main text. The permutation envelope is centred on zero and tests whether any heterogeneity exists; the bootstrap bands are centred on the observed Δ_τ and indicate the precision with which the magnitude of the heterogeneity effect is estimated. In practice, the substantive conclusions are consistent across the two approaches: splitters for which the observed path exits the permutation envelope in the main-text figures are also the splitters for which the bootstrap bands exclude zero at comparable horizons. For completeness, we report the corresponding bootstrap figures below.

by a pre-treatment fit measure such as the post/pre-RMSPE ratio. In the ADH framework the unit of permutation is the treatment label itself: synthetic controls are re-estimated for each placebo unit, and units with poor pre-treatment fit can generate spuriously large post-treatment gaps. The RMSPE ratio corrects for this heterogeneous fit quality. In our setting, by contrast, the episode-level gap paths $G_{i\tau}$ are held fixed across all permutation replications—only the high/low institutional labels are reshuffled. Because no synthetic controls are re-estimated, the pre-treatment fit of each episode is invariant to the permutation, and the null distribution of Δ_τ^π already reflects any heterogeneity in episode-level fit quality through the random allocation of noisier episodes across pseudo-groups.

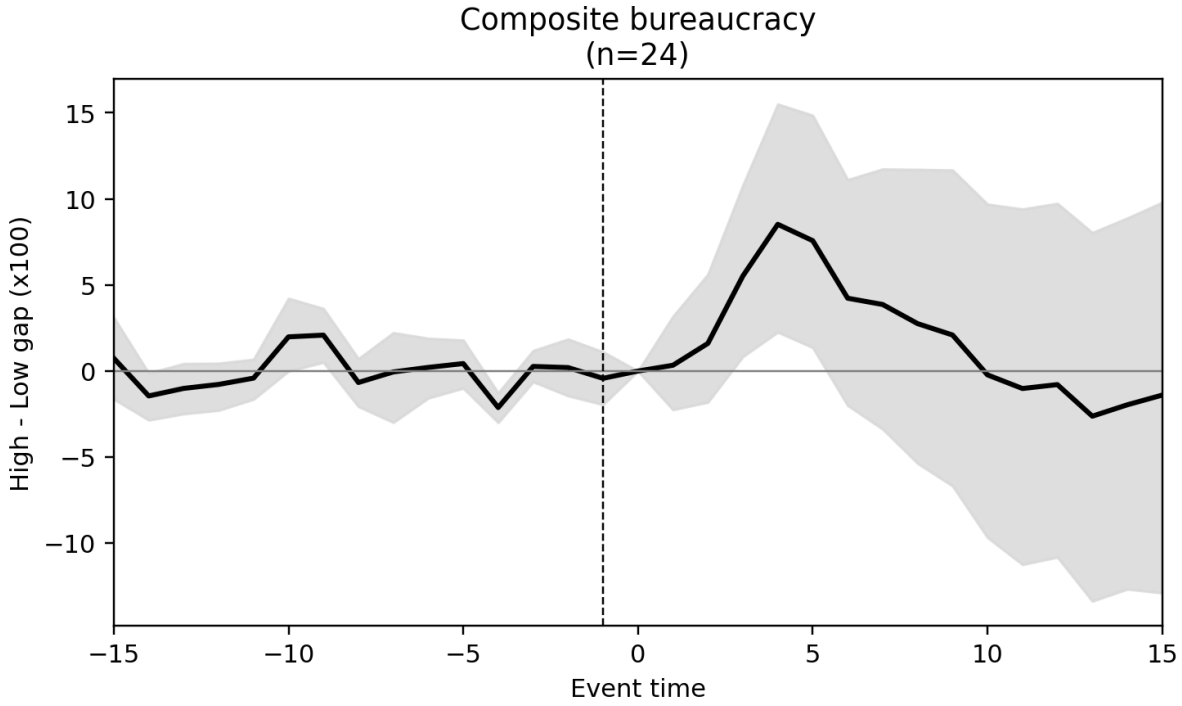


Figure 11: Baseline heterogeneity: composite bureaucracy measure splitters. Shaded regions show pointwise 90% bootstrap confidence bands, constructed by resampling episodes with replacement within the high and low groups. Unlike the permutation envelopes in the main text, these bands are centred on the observed Δ_τ and indicate estimation precision rather than testing the null of no heterogeneity.

G SOTS populism scoring with Qwen

We implemented a full-corpus LLM scoring pipeline for the United States State of the State (SOTS) corpus using `qwen/qwen3-235b-a22b-thinking-2507`. We have 3,355 gubernatorial speeches from all 50 states spanning 1866–2023.

The scoring prompt is from Tamaki et al. (2025). Rather than using a short prompt, we use their long, structured holistic-grading prompt. Each model call used 29 prompt messages: a system instruction, a theoretical definition of populism as an ideational discourse, an explanation of holistic grading, a detailed rubric with six populist and six non-populist categories, ten anchored example speeches ranging from clearly non-populist to highly populist,following Tamaki et al. (2025). We add a SOTS-specific genre instruction for it to be context aware, and finally the target speech. The SOTS instruction was important because these texts are formal, policy-heavy gubernatorial addresses and often contain transcript noise such as headings, page numbers, legislative scaffolding, or other session metadata. The model was therefore explicitly told to score the substantive speech content.

For each speech, we stored not only the scalar LLM score but also the model's overall reasoning, the parsed structured JSON, the raw model output, token counts, finish reason, runtime metadata, model name, temperature (which is always set to 0), whether SOTS context was used, and the number of prompt messages. This made the pipeline auditable and restartable rather than a one-shot CSV export.

We had 12 speeches that were not graded in the first attempt. These nulls came from three different failure modes. First, eight cases were not substantive scoring failures at all: Qwen returned complete answers, including usable scores and reasoning, but the returned JSON was malformed and could not be parsed by the JSON extractor. These eight cases were hand-

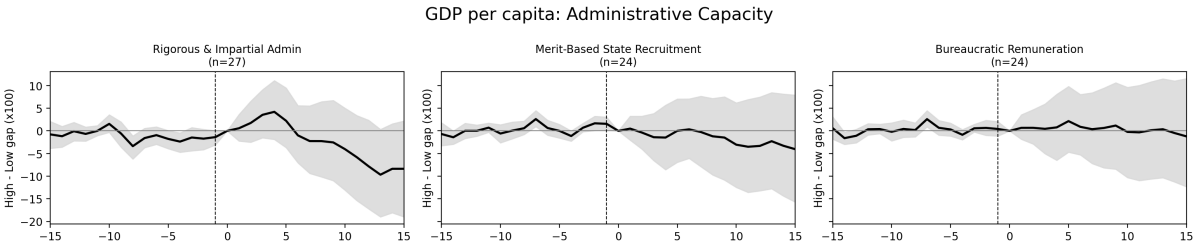


Figure 12: Baseline heterogeneity: broad institutional splitters. Shaded regions show pointwise 90% bootstrap confidence bands, constructed by resampling episodes with replacement within the high and low groups. Unlike the permutation envelopes in the main text, these bands are centred on the observed Δ_τ and indicate estimation precision rather than testing the null of no heterogeneity.

patched. The hand-patched speeches were Arnold Schwarzenegger (California, 2006), George Dekle Busbee Sr. (Georgia, 1979), Eric Holcomb (Indiana, 2017), Robert D. Ray (Iowa, 1982), Thomas Lee Judge (Montana, 1980), Jon Corzine (New Jersey, 2008), John S. Battle (Virginia, 1954), and Mark Gordon (Wyoming, 2020). All eight recovered scores were `0.0`.

Second, two speeches were genuine length failures: the model exhausted the output budget before returning a parseable JSON object. These were B. B. Comer (Alabama, 1911) and James R. Thompson Jr. (Illinois, 1989), both unusually long transcripts. Third, two additional speeches failed at the API/output level despite being normal-length texts: Charles C. Stevenson (Nevada, 1889) and Lincoln Chafee (Rhode Island, 2012). These were re run and successfully got scored.

G.1 Validation

We conduct two validation exercises for the LLM-based populism scores: a comparison against hand-coded human ratings and a cross-model comparison against an independent LLM.

Human validation sample. The human validation sample comprises 100 speeches. The sample is not a simple random draw from the corpus of 3,355 speeches. Instead, it was constructed as a stratified quota sample across four Qwen score bins: 55 speeches with `llm_grade` $= 0$; 20 with $0 <$ `llm_grade` ≤ 0.1; 15 with $0.1 <$ `llm_grade` ≤ 0.2; and 10 with `llm_grade` > 0.2. This design intentionally oversamples non-zero and higher-scoring speeches relative to their share in the full corpus, where 88.3% of speeches receive a score of exactly zero. The human validation sample has a corresponding zero share of 55%.

Table 25 reports the results. Across the 100 speeches, Qwen scores are positively correlated with human scores: the Pearson correlation is 0.420 and the Spearman rank correlation is 0.404. The mean absolute error is 0.105 and the root mean squared error is 0.217; the mean signed difference (Qwen minus human) is 0.013, indicating a slight upward bias. Exact agreement is 55%; 74% of scores are within 0.1 points of the human rating, 90% within 0.2 points, and 94% within 0.3 points.

Because the score distribution is heavily zero-inflated, we supplement these continuous metrics with zero-aware statistics. Of the 100 speeches, 48 receive a score of zero from both Qwen and the human coder. In 7 cases Qwen assigns zero while the human assigns a positive score; in 23 cases the reverse holds. Both assign positive scores in 22 cases. This pattern indicates that Qwen is somewhat prone to assigning small positive scores where a human coder would assign zero, though the frequency of false positives at the zero boundary is modest. For a binary "any populism" classification (score > 0), Qwen achieves precision of 0.489, recall of 0.759, and accuracy of 0.700.

Alternative-model comparison. As a second validation exercise, we re-score the full corpus

with an alternative large language model, GPT-5.5, using the identical holistic-grading prompt and pipeline, and compare it both to Qwen and to the hand-coded sample (Table 25). The two models' score distributions differ substantially: GPT-5.5 assigns a positive score to 39.4% of speeches, against 11.7% for Qwen, so it detects populist rhetoric far more liberally. They are nonetheless correlated — across the full corpus they agree exactly on 61.5% of speeches, fall within 0.1 points on 80.0%, and have a Pearson correlation of 0.633 — indicating that they capture a common signal even though they are not interchangeable. On the human-coded sample, Qwen tracks the human ratings more closely on almost every metric: it has a much smaller mean signed bias (0.013 vs. 0.123), lower mean absolute error (0.105 vs. 0.181) and root mean squared error (0.217 vs. 0.312), higher exact agreement (55% vs. 39%), and more scores within 0.1 (74% vs. 56%) and 0.2 (90% vs. 77%) of the human rating. GPT-5.5's linear (Pearson) correlation with human scores is marginally higher (0.448 vs. 0.420), but its rank (Spearman) correlation is lower (0.360 vs. 0.404). The gap is driven by GPT-5.5's tendency to over-score: it labels 37 of the validation speeches as populist where the human coder assigned zero (against 23 for Qwen), which depresses its precision (0.383) and accuracy (0.570) in the binary "any populism" classification. We therefore retain Qwen as our primary measure and use GPT-5.5 as an external-model robustness check.

Table 25: Human and external model validation of LLM populism scores

	Qwen	GPT 5.5
Panel A Full corpus score distributions		
Full corpus size	3,355	3,355
Exact zero scores	2,963	2,033
Share exact zero (%)	88.3	60.6
Positive scores	392	1,322
Share positive (%)	11.7	39.4
Panel B Continuous agreement with human scores ($N = 100$)		
Mean model score	0.105	0.215
Mean human score	0.092	0.092
Mean model minus human	0.013	0.123
Mean absolute error	0.105	0.181
Root mean squared error	0.217	0.312
Pearson correlation	0.420	0.448
Spearman correlation	0.404	0.360
Exact agreement (%)	55.0	39.0
Within 0.1 points (%)	74.0	56.0
Within 0.2 points (%)	90.0	77.0
Within 0.3 points (%)	94.0	85.0
Panel C Zero aware comparison with human scores		
Both model and human zero	48	34
Model zero, human positive	7	6
Model positive, human zero	23	37
Both model and human positive	22	23
Panel D Binary "any populism" classification (score > 0)		
Precision	0.489	0.383
Recall	0.759	0.793
Accuracy	0.700	0.570
F1 score	0.595	0.517
Panel E Qwen and GPT agreement on full corpus ($N = 3{,}355$)		
Exact agreement (%)	61.5	
Within 0.1 points (%)	80.0	
Within 0.2 points (%)	91.9	
Within 0.3 points (%)	95.3	
Pearson correlation	0.633	
Spearman correlation	0.445	

Notes: Panels B to D use a human validation sample of 100 speeches hand coded by the authors and matched to the Qwen and GPT 5.5 scoring files. The sample is stratified by Qwen score bin and intentionally oversamples nonzero speeches relative to the full corpus. All statistics are unweighted. "Within x points" reports the share of speeches for which the absolute difference between the model score and the human score is below x. The binary classification in Panel D defines a positive case as any score greater than zero. The main empirical treatment is stricter: a governor is classified as high populism only when the governor's mean Qwen score exceeds the 90th percentile of the 1866-2023 state year distribution, with cutoff equal to 0.06. GPT 5.5 is used as an external model validation exercise; all main results use Qwen scores.

G.2 Construction of the state-year panel

We construct the final state-year panel in two stages.

Stage 1: Governor-level aggregation. We begin by aggregating speech-level LLM populism scores to the governor level. Speech scores are matched to the SOTS metadata using the speech identifier, after which governor names are standardized to reduce false duplication. Standardization includes reordering names recorded in last-first format, consolidating within-state spelling variants where string similarity and overlapping years indicate the same individual, and manually correcting remaining typographical inconsistencies. We then collapse speeches by `(state, governor)` to obtain the governor-level mean populism score, as well as the number of speeches and the observed year range. Our main measure, `llm_pop_mean`, is defined as the average LLM populism score across all speeches attributed to a given governor.

Stage 2: Mapping to a panel. We map these governor-level measures onto a state-year panel. For each state we enumerate all years from the first observed speech year through 2023 and assign a governor to each state-year using a multi-source roster procedure. This procedure combines an official governor roster—with the panel skeleton drawn from Kaplan (2021)—with manual corrections, post-2020 updates, speech-based fallback assignments, and term-boundary adjustments in cases where roster records and speech evidence conflict. Because governor names in the roster and speech corpus do not always coincide exactly, roster entries are fuzzy-matched to the cleaned governor identifiers prior to merging. Each state-year is then assigned the populism profile of the corresponding governor; state-years for which no governor-level LLM score can be matched are retained with missing populism measures.

Transition years. For years in which a gubernatorial transition occurs, we assign the state-year observation to the incoming governor. This rule reflects the timing of State of the State addresses, which are typically delivered early in the legislative session—often in January or February—so that the incoming governor governs for the larger share of the calendar year. Even if the outgoing governor delivered a speech early in the transition year, the panel row is coded to the incoming officeholder's tenure. Importantly, this rule operates only at the panel-construction stage: each speech remains attached to the governor who actually delivered it when computing governor-level average populism scores. We flag such cases with the indicator `is_transition_year`, which identifies state-years in which the speech data contain more than one distinct governor.

Populism classification. Our main populism variable is a binary indicator derived from the governor-level mean LLM populism score. The variable *Pop* equals 1 when `llm_pop_mean` exceeds the 90th percentile of the 1866-2023 state-year distribution, and 0 otherwise; observations with missing LLM scores remain missing. In our sample the 90th-percentile cutoff is 0.06.[22]

H Further detail on data construction

H.1 Party coding

Governor party affiliation is coded as a set of dummy variables, with blank entries left missing. Democratic party labels—including `democrat`, `democratic`, `democratic-farmer-labor`, and `silver-democratic`—are coded as 0. Republican party labels—including `republican` and

[22] As robustness checks we also construct Pop_{75}, coded as 1 when `llm_pop_mean` exceeds the 75th percentile (cutoff: 0.02), and 0 otherwise. Using a continuous specification or a simple sign-based indicator, $Pop_{positive}$ [where $Pop_{positive} = 1$ if `llm_pop_mean` > 0 yields similar results.

`independent-republican`—are coded as 1. All other non-empty party labels are coded as 2, denoting Other.

H.2 Bureaucracy

We include two state-level institutional indicators derived from the dataset of Ash et al. (2022). The first captures whether a state had enacted a civil-service reform law by a given year; the second captures whether a state had adopted a formal merit system by that year. Substantively, these variables proxy the extent to which public employment was insulated from patronage and organized according to professionalized administrative rules. In the panel they should be interpreted as adoption indicators: once a state enacts the relevant reform, all subsequent years are coded as exposed to that institutional arrangement.[23][24] We complement these rule-based indicators with a measure of bureaucratic scale drawn from a state-year panel of U.S. Census Bureau public employment and payroll data that we construct for 1944–2024.

Sources and coverage. The panel combines modern Annual Survey of Public Employment and Payroll (ASPEP) and Census of Governments extracts for 1992–2024 with Census Bureau historical series reaching back to 1944. A single harmonized set of variable names and function definitions is maintained across the two sources. Because the Census Bureau did not conduct the survey in 1996, that year is absent. The resulting panel is unbalanced, since some series are not published or consistently recorded in earlier years.

Scope. The unit of observation is a state-year. We restrict attention to *state government*—excluding local government and combined state-local aggregates—to align with our focus on state administrative capacity. Within state government we retain four functional categories that are consistently available: (i) total state government, (ii) financial administration, (iii) other government administration (labeled "central administration" in some historical vintages), and (iv) judicial and legal.

Variables. For each function we construct a common set of measures: full-time employees (headcount), full-time equivalent employment (FTE), full-time payroll, part-time employees, and part-time payroll. When part-time components are not directly reported in the archival data, we derive them as residuals (e.g., total employees minus full-time employees). Part-time hours are available only in the modern ASPEP extracts and are therefore missing before 1992 by construction. All payroll figures are recorded in nominal dollars. Special codes used in the historical database to flag unpublished items (e.g., `-11111`) are recoded to missing; zero values are retained as recorded.

Harmonization. We harmonize function labels across vintages and standardize the state identifier to a single time-invariant FIPS code. The resulting panel merges directly into our main empirical dataset. Our primary measure of bureaucratic scale is total state-government headcount—combining full-time and part-time employees. Conceptually, it captures the size of the state's administrative apparatus: whereas civil-service reform measures the *rules* governing public employment, this variable measures the *organizational scale* of the bureaucracy itself.

[23] These measures do not necessarily imply that the reform remained fully in force in every later year, since some states subsequently weakened or partially repealed such arrangements. They therefore capture institutional adoption rather than a precisely measured contemporaneous level of administrative professionalization.

[24] The civil-service reform year is recorded as missing for Texas in Ash et al. (2022). Because Texas never enacted a civil-service reform law, we code the indicator as 0 for all Texas state-years.

H.3 Income

Annual per-capita personal income series are drawn from the Federal Reserve Economic Data (FRED) database and are recorded in nominal terms. Coverage generally begins in 1929 for most states, although Alaska, Hawaii, and occasionally other later-added entities start in subsequent years depending on the exact series.

H.4 Missing data

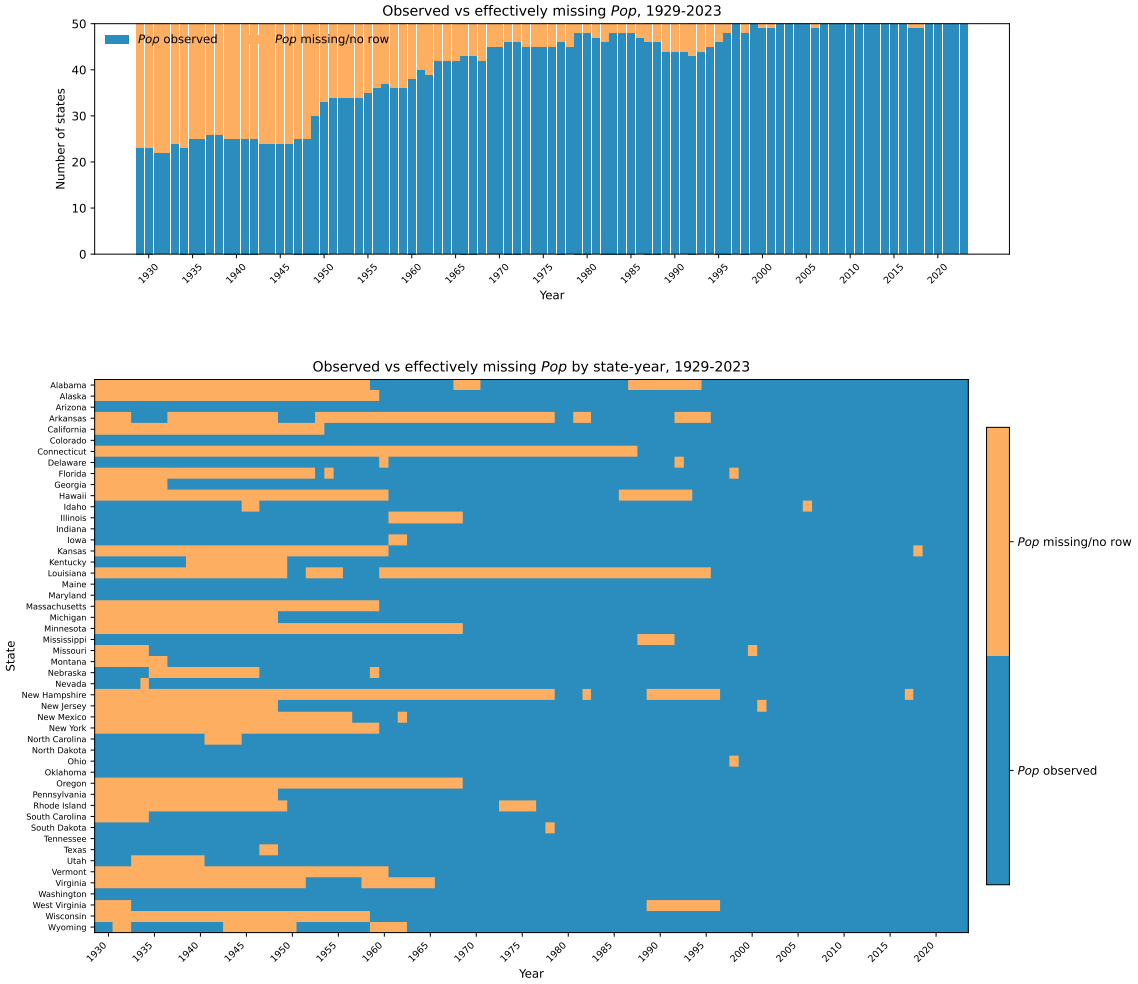


Figure 13: Populism score availability, 1929–2023. The upper panel plots the number of states with an observed populism indicator *Pop* by year. The lower panel shows coverage at the state-year level; orange cells indicate state-years where *Pop* is unavailable, either because the state-year row is absent from the speech panel or because *Pop* is missing despite the row existing. Of the 4,750 state-years in the 1929–2023 target grid, *Pop* is observed for 3,861 (81.3%).

The working LLM panel contains 4,691 state-year observations spanning 1866–2023 and 155 variables.

Populism scores. The governor-level LLM populism score is missing in 296 observations (6.3%), while the dictionary-based populism score is missing in 302 observations (6.4%). These gaps are concentrated in state-years for which the roster governor could not be matched to a speech-based governor score rather than in any single time period or region. The largest concentrations occur in Alabama (55 state-years), Arkansas (44), Louisiana (40), Utah (28), Wyoming (17), and Nebraska (15); all 50 states retain at least some non-missing populism data. There are no cases in which the dictionary score is present but the LLM score is missing. The asymmetry runs only in the other direction: six state-years have an LLM score but no dictionary score (South Dakota 1907–1908 and Texas 1911–1914). The within-governor LLM variance measure is missing more often (674 observations, 14.4%), but most of this is mechanical: 378 rows correspond to governors with only a single speech, for whom within-governor variance is undefined, while the remaining 296 are the same rows in which the LLM score itself is missing.

Macroeconomic variables. The macroeconomic variables display window-based missingness rather than sporadic holes. Personal income and per-capita income are complete from 1929 onward; unemployment from 1976 onward; nonfarm employment from 1990 onward; and state GDP from 1997 onward. Real per-capita income is available only from 2008 onward. The Akcigit innovation and tax variables are complete within their observed window (1937–2010) and missing outside it.

Bureaucracy variables. The bureaucracy variables are mostly missing before the post-war period. Total state-government headcount is observed from 1946 onward and full-time-equivalent employment from 1952 onward, with a few within-window gaps concentrated in 1951, 1958–1960, and 1996. Bureaucracy subcomponents—especially the judicial/legal and payroll series—are substantially sparser than the total-employment measures. The institutional reform dates from Ash et al. (2022) and Vannoni et al are unevenly populated. The civil-service reform year is nearly complete, missing only for Texas (which we code as never having adopted the reform; see footnote above). The civil-service repeal date is missing for most observations, as expected, since repeal or rollback is recorded only for the subset of states that reversed earlier reforms.

Conditional-language and other policy variables. The conditional-language variables are the sparsest policy-content measures. They are unavailable before 1960, and even within 1960–2012 their coverage follows a clear biennial pattern: data are effectively concentrated in even years, whereas odd years from 1961 to 2011 are missing for nearly all states. By the late period odd-year coverage is entirely absent. Among even years, 2008 and 2010 each have one missing observation, and 2012 has 13. The professionalized-employment indicator is available only for 1965–1984 and covers 47 states.

H.5 Correlates of states

We construct four controls from Version 2.6 of the Correlates of State Policy Project (CSPP), which provides a state-year panel with common identifiers including state FIPS code and year (Grossmann et al., 2021). Specifically, we extract `percentBlack`, `percentForeignBorn`, `firms`, and `bankrupt` from the CSPP file and keep only the state FIPS identifier and calendar year needed to merge these variables to the main panel.

The CSPP codebook defines `percentBlack` as the percentage of the population identified as Black by the Census race variable, and `percentForeignBorn` as the percentage of the pop-

ulation identified as foreign-born by the Census (Bullock, 2020). The codebook defines `firms` as the number of firms in a state, where a firm is a business organization consisting of one or more domestic establishments in the same state and industry under common ownership or control; this series is sourced from the U.S. Census Bureau's Statistics of U.S. Businesses (U.S. Census Bureau, 2012). We merge these CSPP variables to the main panel using the key (`statefips`, `year`) after harmonizing the CSPP FIPS identifier to the naming convention used in the main dataset.

Because these series are not observed annually for all state-year cells, we construct filled versions within state. For the proportion variables `percentBlack` and `percentForeignBorn`, we first transform the observed series to the logit scale,

$$z_{st} = \log\left(\frac{x_{st}}{1 - x_{st}}\right),$$

after moving any boundary values slightly inside the unit interval. We then linearly interpolate missing interior years and linearly extrapolate endpoint years within each state on this transformed scale. The completed series are mapped back to levels using the inverse-logit transformation,

$$x_{st}^{\text{fill}} = \frac{\exp(z_{st}^{\text{fill}})}{1 + \exp(z_{st}^{\text{fill}})}.$$

For the nonnegative variables `firms` and `bankrupt`, we instead work on the log scale. Let

$$z_{st} = \log(x_{st}),$$

after replacing any nonpositive value by a very small positive constant. We then linearly interpolate and extrapolate z_{st} within state over time and recover the completed level series by exponentiation:

$$x_{st}^{\text{fill}} = \exp\left(z_{st}^{\text{fill}}\right).$$

The final filled variables are therefore `percentBlack_fill`, `percentForeignBorn_fill`, `firms_fill`, and `bankrupt_fill`. In the empirical specifications, we use the logarithms of these completed series. For variables that are strictly positive after the filling step, we define

$$\texttt{ln\_percentForeignBorn\_fill}_{st} = \log(\texttt{percentForeignBorn\_fill}_{st}), \quad \texttt{ln\_firms\_fill}_{st} = \log(\texttt{firms\_fill}_{st})$$

Because a small number of observations in `percentBlack_fill` are exactly zero, we use a small offset when taking logs:

$$\texttt{ln\_percentBlack\_fill}_{st} = \log(\texttt{percentBlack\_fill}_{st} + 0.01).$$

Thus the controls used in the regressions are `ln_percentBlack_fill`, `ln_percentForeignBorn_fill`, `ln_firms_fill`, and `ln_bankrupt_fill`.

The raw control variables from the Correlates of State Policy Project are highly sparse in annual state-year data. In the civil service reform estimation sample, for example, `percentBlack` and `percentForeignBorn` are each missing for 3,509 of 3,861 observations, `firms` is missing for 3,159 observations, and `bankrupt` is missing for 3,521 observations. After constructing filled annual series by interpolation and extrapolation within state, the corresponding logged controls are fully observed in the estimation sample. Thus, it is the filling procedure rather than the log transformation that resolves the missing-data problem.