For Online Publication

A Additional institutional background

A.1 Other forms of policy experimentation

While we focus in this paper on the form of policy experimentation through experimentation points, it is important to note that policy learning in China also takes place in several other forms that may not squarely fit into the conventional definitions of policy experimentation (Heilmann 2008b).

Specifically, there are three such forms of policy learning. First, "interim policies" (*Shixing/Zanxing*). These are provisional policies with clear expiration dates, but they typically apply to the whole country and do not have regional variation. This approach is often used to figure out implementational logistics of a policy before finalizing them in the national legal documents, rather than to learn about the cost and benefit of the policy itself. Second, "demonstrational zones" (*Shifanqu*). These are regions selected as "positive examples" in implementing certain policies, which the central government encourages the rest of the country to emulate. The main purpose of setting up these zones is not to learn about the policy, but to promote the diffusion of a new policy among the local governments. Third, a number of policy experiments target firms (rather than a specific region). The main purpose of such experiments is often to guide the reform of state-owned enterprises.

A.2 Background of four policy experimentation examples

A.2.1 Carbon emission trading

In October 2011, the National Development and Reform Commission designated seven regions to participate in the pilot of carbon emission trading, including Beijing, Chongqing, Guangdong, Hubei, Shanghai, Shenzhen and Tianjin. These experimentation sites were required to design and set up their own carbon markets, following certain general guidelines provided by the central government. Specifically, the experimentation sites had the discretion to determine details like the coverage of the local carbon market, the emission target, and the allowance allocation, etc. Different from the traditional "cap and trade" system, China's carbon markets all followed a less stringent "tradeable performance standard" system, where the regulator sets benchmarks for carbon emissions per unit of output and allows emitters to trade allowances (Cui, Zhang, and Zheng 2021).

The seven pilot carbon markets started operating in 2013, with carbon allowances varying from 30 MT in Shenzhen to 338 MT in Guangdong, and emission coverage varying from 33% in Hubei to 60% in Tianjin. Despite being riddled with controversy regarding its effectiveness, activeness, and economic impacts, the carbon emission trading system was rolled out to the whole country in 2021, after China announced its carbon neutrality plan.

A.2.2 Separation of permits and business licenses

In order to simplify the administrative process of starting a business, the Chinese government started a policy experiment on separating permits and business licenses. With the combination of multiple business credentials, enterprises are able to conduct regular business operations by virtue of the business license alone, instead of applying for permits from different government branches. Starting in Shanghai in 2015, the experimentation was coordinated by the Ministry of Commerce. More prefectures were included in the second wave of experimentation in 2017. A year later, separation between the business permit and license was carried out on the first lot of 106 administrative approval items for enterprises nationwide.[1] The government continued to experiment with this policy after that, aiming at expanding the scope of the policy to more items requiring administrative approval.

A.2.3 Agricultural catastrophe insurance

Featuring high payout ratio but low market demand in terms of risk perception, the agricultural insurance in rural areas has had relatively low participation rate. Starting in 2017, the ministry of agriculture started piloting for catastrophe insurance that features premium subsidies, creating stronger incentives for farmers to voluntarily participate in the program. The first round of experimentation explicitly targets 14 provinces, initially covering farmers of basic grains and selected oil crops and livestock. The list of insured risks was extended in 2019. Until 2021, the government hasn't yet explicitly rolled out the policy to the entire country. Despite the extended list of insurers, increased liability and coverage, some argue that the lack of critical data, under-developed technique, and the lack of awareness in most rural areas still stand in the way of fostering rural resilience (Yu and Yu 2020).

A.2.4 Fiscal empowerment reform

In the Chinese administrative hierarchy, each province administers several prefectural cities, and each prefectural city administers a number of counties. Many have argued that when prefectural cities have fiscal control over counties, the lack of fiscal autonomy of rural counties would hinder their economic development (Wang 2016; Bo 2020). To address this issue and to foster county economic growth, in 2003, the central government started a large-scale policy experimentation on county fiscal empowerment reform. As illustrated in Appendix Figure A.21, the reform primarily empowers counties by flattening the government hierarchy: before the reform, prefectural cities have fiscal controls over counties, while after the reform, counties can bypass the prefectural government and directly respond to the provincial government. Within a decade, more than 1,100 counties in China were assigned as the experimentation sites of the reform. The experimentation was rolled out in multiple waves. Based on the central government's document that

[1] See `http://english.www.gov.cn/policies/latest_releases/2018/10/10/content_281476339291118.htm` for details

guides the fiscal empowerment reform, we collect information on the timing at which participating experimentation sites began the fiscal reform.

As summarized in Li, Lu, and Wang (2016), the existing literature studying the county fiscal empowerment reform reports mixed findings on its effectiveness in promoting local GDP growth, which is highly sensitive to the sample period being used for the analysis. Such mixed findings in the literature could be attributed to the fact that the reform has heterogeneous impact on localities with different economic conditions, and there exists large differences in the underlying site selections throughout the experimentation.

A.3 Government organizational reform

We use the context of China's government organizational reform to understand the organizational environment under which policy experimentation take place.

Since 1998, China has been conducting a series of vertical management (*Chuizhi Guanli*) reforms. Such reforms essentially switch central government ministries and commissions from multi-divisional form (M-form) to unitary form (U-form), by shifting the administration of local bureaus in terms of their personnel, finance, and facilities from the local governments to the corresponding central ministry or commission. For example, before 1999, local securities regulatory bureaus were under the jurisdiction of provincial governments (M-form). After the vertical management was implemented in the security regulatory bureaus in 1999, they came under the direct administration of the central government's Securities Regulatory Commission (U-form).

The literature on organizational theory distinguishes between two types of organizational structure (Chandler 1962; Williamson 1975): multi-divisional form (M-form), which consists of self-contained units in which complementary tasks are grouped together; and unitary form (U-form), which consists of specialized units in which substitutable or similar tasks are grouped together (see Appendix Figure A.22 for an illustration of the distinction between M-form and U-form organizations). While the U-form organizational structure can better take advantage of the economies of scale, the M-form structure provides more flexibility for experimentation. Under the M-form, local managers are able to ensure attribute matching across multiple dimensions, which makes it easier to carry out local experimentation. In contrast, under the U-form, inter-organizational coordination is needed to achieve attribute matching, which complicates potential experimentation (Qian, Roland, and Xu 2006).

The vertical management reforms took place in a staggered fashion over an extended period of more than two decades. See Appendix Table A.13 for a list of the ministries that underwent the vertical management reforms and the years at which they took place.

B Auxiliary data sources

We match our dataset on policy experimentations with several additional sources of data, which we describe in detail below.

B.1 Biographical information of politicians

We collect detailed biographical information on the universe of Chinese central ministers and local (provincial and prefectural) leaders during our four-decade sample period. For each politician in our sample, we have information on his hometown, date of birth, level of education, current job title, past work history, etc.

Following Wang, Zhang, and Zhou (2020), we estimate each politician's *ex ante* promotion prospect in each year, which is a flexible function of his age and official rank in the bureaucratic system, and can be used as a proxy for his career advancing incentives.

Specifically, we estimate each prefectural city leader's *ex ante* likelihood of promotion in each year, as a flexible function of his age when starting the term/position, position and official rank in the bureaucratic system. Our data documents observations across 4,980 terms of office, in 333 prefectural cities in China from 1985 to 2017. At the politician level, we document his age, educational background, current hierarchical level in the government, previous work experience and promotion status after the term.

As described in Wang, Zhang, and Zhou (2020), mandatory retirement age varies with the hierarchical ranking of a city leader, so both the age and hierarchical level of city leaders at the start of their office term largely determine their likelihood of promotion. We therefore estimate the effects of initial age and hierarchical rank at the start of office (start age and start level, respectively, and their interaction term) on promotion likelihood.

Specifically, we use a Probit model with the estimated coefficients to construct the career incentive index as follows:

$$\hat{y}_{pt} = \Phi^{-1}\left\{\hat{\alpha} \cdot startage_{pt} + \hat{\beta} \cdot level_{pt} + \hat{\gamma} \cdot startage_{pt} \times level_{pt}\right\}. \quad (6)$$

Note that t here stands for term of office. The observational level is prefecture by term, so the career incentive index we constructed will be a fixed value throughout a given term of office. Appendix Table A.14 shows the estimated coefficients in the first stage. The first two columns shows estimates by LPM and column 3 and 4 shows estimates by Probit. The sign and magnitude of the estimated coefficients are consistent with Table 2 from Wang, Zhang, and Zhou (2020).

B.2 Government organizational structure

We collect information on the organizational structure of all government ministries and commissions in China in the past four decades. Following the definition of Qian, Roland, and Xu (2006), we categorize each central ministry/commission as either an M-form organization or a U-form one. Some central ministries and commissions, such as the ministry of foreign affairs, only operate at the national level and do not have local branches, and are therefore not applicable to the M-form/U-form distinction.

We also collect detailed information on government organizational reforms in China during our sample period, which enables us to identify ten cases in which an M-form ministry/commission switches into U-form after a certain year. The panel is unbalanced due to ministry cancellations and mergers during this period. For ministries that merged with each other, the unit of analysis is the eventually merged ministry throughout the sample period.

B.3 Local socioeconomic conditions

We collect comprehensive panel data on regional socioeconomic conditions from the annual statistical and economic yearbooks published by the national bureau of statistics, which covers all the provinces, prefectural cities, and counties in China between 1993 and 2018. The data contains detailed information on economic growth, demographics, and public good provision, and can be matched to the experimentation point status assigned by each round of the policy experiments.

B.4 Local fiscal expenditure

We collect county-level fiscal revenue and expenditure data from the National Prefecture and County Finance Statistics Yearbooks between 1993 and 2006. The dataset covers all counties in China, and provides detailed yearly information on fiscal revenue and expenditure by each domain. Over our 14-year sample period, the definitions of the fiscal expenditure domains changed several times, but six broadly defined domains remained consistently reported every year: general administrative cost, infrastructure, economic production, agriculture/forestry/fishing, science/culture/education/medicare, and others. We thus focus on these six domains, and match every policy experiment during this period to its most relevant fiscal domain.

B.5 Land revenue of the local government

We measure land revenue received by the local government, particularly those driven by the amount of land suitable for real estate and commercial properties development and local demand shocks. We use the interaction of both as an instrumental variable for the land revenue income of local government, following Chen and Kung (2016).

We match land revenue data (based on Fiscal Statistical Compendium for All Prefectures and Counties, from which data is available for the period 19992006, and the website of the Land Transaction Monitoring System, http://www.landchina.com, for 2007-2008 data) with geographic elevation data from United States Geographic Service (USGS) Digital Elevation Model (DEM) at 90-meter resolution, which allows us to estimate the percentage of land unsuitable for real estate development. Moreover, we match the land revenue data with the housing price data from the *Statistical Yearbook of Regional Economics* (2000-2009), which proxies for land demand. We used the interaction of both as an instrumental variable for the land revenue income of local government. The construction of such instrumental variable follows essentially that of Chen and Kung (2016).

B.6 Five Year Plans

We collected all the documents from the Five Year Plans issued by the State Ministry and all its branches, which normally contain detailed economic development guidelines as well as targets for all its regions. When a policy experimentation is mentioned in one of the Five Year Plans, the central government demonstrated solid resolution to promote the idea of the policy and track progress of its implementation.

B.7 Local political and social unrest

We compile data on episodes of political and social unrest throughout China from the Global Database of Events, Language, and Tone (GDELT), one of the largest databases on global political events. See www.gdeltproject.org for details of the GDELT Project.

C Organizational structure and experimentation tendency

While many factors could contribute to the patterns of the number of policy experiments initiated over time, we next explore a particular set of factors related to the organizational structures of the political bureaucracy and the compatibility of different structures with the ability to coordinate and implement complex policy experimentation.

Theories in organizational economics distinguish between two particular types of organizations that may have first-order implications for the ability of the organizations to coordinate experimentation. The multi-divisional form (or M-form) organizations consist of self-contained units in which complementary tasks are grouped together. In the context of political organizations, a typical M-form structure entails that local, say provincial government, has jurisdiction over its own bureau of finance, bureau of labor, bureau of agriculture, and bureau of education, etc. As a result, each provincial government can function as a standalone unit and coordinate policies and tasks across bureaus within the localities without necessarily the need to coordinate with other localities. In contrast, the unitary form (or U-form) organizations are decomposed into specialized units in which substitutable or similar tasks are grouped together. In the context of political organizations, a typical U-form structure entails that central government has jurisdiction over the ministry of finance as well as its local bureaus in each province, for example. As a result, policies related to finance can have a streamlined procedure for implementation as the national finance ministry can directly coordinate its local counterparts in each locality. In other words, the M-form organizations are more decentralized and flatter, while the U-form organizations are centralized and vertical.

M-form and U-form organizations represent an organizational trade-off between flexibility and efficiency. Under the M-form structure, local managers are able to ensure attribute matching across multiple dimensions, making it substantially easier to carry out small-scale yet complex experiments that may involve coordination across several arms of the government. On the other hand, under the U-form structure, inter-unit coordination is needed to achieve effective attribute matching, which complicates and hinders small-scale experiments. However, the U-form organizations benefit from potential economies of scale: policies are easy to scale up to the entire country under U-form organizations, and standard decision-making can ensure that the same, compatible policies in a particular domain are implemented throughout the country.

Accordingly, one often observes M-form organization structure in government bureaucracy for small government or government at earlier stage of the development, and U-form organization for developed polities where gains from economies of scale may outweigh flexibility. As described in Section A.3, the Chinese government has undergone a series of restructures of its organizations, moving away from M-form to U-form across many ministries and government commissions, and shifting the control over the ministries' personnel, funding, and property rights from the local governments to the upper-level ministerial units.

We formally examine whether the M-form organizations in government bureaucracy are better at facilitating policy experimentation, and U-form organizations are relatively worse at coordinating and initiating such experiments. In particular, we identify the impact of a M-form to U-form transition on the number of policy experiments initiated by the ministry or commission. Following an event study design, we estimate the following specification:

$$y_{mt} = \sum_k D_{mt}^k \cdot \beta_k + \delta_m + \theta_t + \varepsilon_{mt}, \tag{7}$$

where y_{mt} is the total number of policy experiments initiated by ministry/commission m in year t, and D_{mt}^k is the years relative to ministry/commission m's switches from M-form to U-form. We include a full set of ministry/commission fixed effects (δ_m), as well as a full set of calendar year fixed effects (θ_t), allowing us to exploit variations within ministry/commission and exploit the fact that different ministries/commissions went through the M- to U-form transition in different years. The baseline specification clusters the standard errors at the ministry/commission level.

Appendix Figure A.23 plots the non-parametrically estimated D_{mt}^k coefficients. Consistent with the theoretical predictions, following the transition to U-form, we find that the vertically managed ministries significantly decrease the amount of policy experimentation they administer. The decrease is substantial in magnitude, representing a 59.4% reduction in the number of policy experimentation initiated over the first three years after the organization restructuring, relative to the average level just prior to the U-form transition. Suggesting a causal interpretation, we do not find any noticeable pre-trend leading up to the U-form transition; in other words, there does not appear to be strategic timing of the U-form transition targeting ministries or departments on particular trajectories in terms of the policy experiments they initiated, neither are there substantial preemptive experiments just prior to the transition away from M-form organization.

Taken together, the results presented above indicate that the flat, decentralized organizational structure provides the flexibility and relative easiness to coordinate, which in turn facilitates policy experimentation. At least part of the decline in the number of experiments in the recent decade that we observe is due to a shift away from the flat, multi-division organizations of the state ministries to a more centralized structure that benefits from the economies of scale, which may be an inevitable outcome as the development reaches a relatively high and mature level. A simple back of the envelope calculation suggests that one could attribute a reduction of five policy experiments per year to the shifts of ministries to U-form. Though importantly, such a shift to U-form organizations that benefit from the economies of scale may push against the *increasing* need for policy experimentation, as reforms and the policy space become more complex and uncertain with the social and economic development.

D Optimal experimental design simulations

In addition to learning about the true underlying treatment effects and persuading other agents who might hold different priors, the central government as a decision maker may carry alternative objectives. If this is the case, then the unrepresentative roll-out of experiments may be justified. We conduct a quantitative exercise to examine that if we incorporate two specific objectives — the central government caring about subjective expected utility from the policy, or about the welfare of the experimentation sites — how much of the positive selection that we observe can be justified.

For the following simulations, we use data from three policy experiments with t-statistics on GDP per capita at the 25th, 50th, and 75th percentiles: (1) "Reform of Comprehensive Administrative Law Enforcement System for Business" (*t-stat* = 0.08), (2) "National Care and Service System for Left-behind Migrant Children in Rural Areas" (*t-stat* = 0.53), (3) "Tax Classification and Coding of Goods and Services" (*t-stat* = 8.52).

D.1 Simulations with ambiguity aversion following Banerjee et al. 2020

Overview First, we examine the incentives of subjective expected utility, in addition to learning and persuasion. Following Banerjee et al. 2020, we simulate the optimal experimentation design, parameterizing the model based on the experimentation setup and estimated heterogeneous treatment effects from Section 6.2. As predicted by Banerjee et al. 2020, when the decision maker (central government) places heavier weight on its subjective expected utility, deterministic experimentation becomes more justified than randomization. However, even if we place 100% of the weight on the decision maker's subjective expected utility, the optimal design of the deterministic experimentation would only induce positive selection with mean *t-stats* = (0.006, 0.051, -0.006) for each of the three experiments, which is substantially lower than the positive selection that actually occurs. Under reasonable assumptions, motivations to maximize subjective expected utility alone is *not* able to justify the level of deviation from representativeness in experimentation site selection that we observe.

Banerjee et al. 2020 present a model wherein a decision maker (DM) must balance maximizing their own subjective expected utility, a function of the DM's priors, against maximizing expected utility for others with potentially hostile priors.

Specifically, the DM aims chooses experimental design ϵ and allocation rule α (a mapping of experimental data to policy decision) to maximize the decision problem (DP):

$$\lambda \mathbb{E}_{h_0,\epsilon}[u(p, \alpha(e, y))] + (1 - \lambda) \min_{h \in H} \mathbb{E}_{h,\epsilon}[u(p, \alpha(e, y))]$$

where H is the set of all relevant priors, h_0 is the DM's own prior, p is a vector of treatment effects conditional on covariates, $\alpha(e, y)$ is the allocation rule dependent on experimental assignment e and outcome data y, $u(p, \alpha)$ is the average treatment effect of the policy, and $\lambda \in [0, 1]$ a parameter controlling how much the DM values their own utility relative to satisfying other priors. Thus, pure subjective utility maximization is the case where $\lambda = 1$.

We simulate the optimal experimental design for each of the three policy experiments with the following procedure:

1. We first compute the vector of treatment effects p for each county that receives treatment, using a difference-in-difference specification with controls for pre-experiment GDP and province fixed effects. There are (49, 946, 138) counties that receive treatment during the first waves for the three experiments. Using these treatment effects, we then impute treatment effects for the non-treated group based on the covariates GDP and province. The total sample consists of 2,010 counties, and the mean treatment effect is an increase in GDP per capita of (6.00%, 17.75%,4.90%) over the pre-period quantity (s.d. = (17.00%,28.88%, 25.90%)).

   Since the number of covariates influencing the outcome must be larger than the size of the treated sample (otherwise, the experiment may be sufficient to characterize the effect of the covariates and perfect information is attained), we split the pre-experiment GDP into 2,010 bins corresponding to the 2,010 counties.

2. Next, we construct the space of priors H. Each prior h_p consists of 10 sub-priors $p_s \in h_p$ which are equally weighted in likelihood. Each sub-prior consists of 2,010 expected treatment effects (one per county) $subprior_{p_s,c} \in h_p \in H$, following the data generation process:

   $$subprior_{p_s,c} = \beta_c + \gamma_{p_s} + \eta_{p_s,c} \qquad \gamma \sim U[-2\bar{\beta}, 2\bar{\beta}], \eta \sim U[-\beta_{max}, \beta_{max}]$$

   where p_s indexes a particular sub-prior, c indexes a county, β is the true treatment effect, $\bar{\beta}$ the mean treatment effect, and β_{max} the largest observed treatment effect. Hence, the sub-prior can be broken into three terms: the true treatment effect β_c, an idiosyncratic bias on the effect of the treatment for each prior γ_p, and random noise $\eta_{p,c}$. Hence, the expected value of each sub-prior's treatment effect is the true treatment effect.[2] We construct 1,000 priors to form H and run the simulation with the DM holding each of these priors as their own (h_0) with the other priors treated as hostile.

3. Then, we construct the space of potential solutions to the DP. A solution to the DP consists of an experimental design ϵ and an allocation rule α. Each experimental design randomly draws counties equal to the number of counties treated under the real experiment for treatment. 1,000 of these experimental assignments are generated in the simulation. The allocation rules take the form

   $$\alpha(e, y) = \mathbb{1}[\bar{y}^1 + \delta > \bar{y}^0]$$

   where $\bar{y}^1, \bar{y}^0$ are the mean outcome for the treated and non-treated groups respectively, and δ is a parameter that can be adjusted to characterize different potential allocation rules. 5 values of $\delta : \{-2\bar{\beta}, -\bar{\beta}, 0, \bar{\beta}, 2\bar{\beta}\}$ are selected to construct 5 allocation rules. Thus, there are 1000 designs X 5 allocation rules = 5,000 random potential solutions to the DP.

[2] We formulate priors as being composed of discrete sub-priors rather than a continuous distribution for computational feasibility.

4. Once the priors and potential DP solutions have been constructed, we proceed to maximize the DP by finding the optimal solution for each prior $h \in H$. We solve eleven versions of the DP for each prior, corresponding to $\lambda \in \{\frac{x}{10} | x \in \{0, 1, \ldots, 10\}\}$. For each of these (deterministic experimental design) solutions, we then compare its expected value to the expected value under the RCT experimental design (where the set of sampled experimental designs is taken as representative of the total), and select whichever is higher as the optimal solution.[3]

5. Once an optimal experimental design has been found for each prior, we compute t-statistics for group balance under the design and store it.

6. For each set of parameters, we repeat steps 1 - 5 for 1000 times total, given that the priors (and treatment effects under the general experiment case) are randomly generated.

The results from these simulations are displayed in Figure A.24. Mean t-statistics are (0.006, 0.051, -0.006) for the Law enforcement for business, Care for left-behind children, and Tax classification policies respectively.

Differential quality of information: Selection of experimentation sites may be influenced by the fact that counties may be differentially capable of running experimental policies, resulting in differential quality of the informational signal arising from selected counties for treatment. Given that richer counties typically have more government capacity and ability to execute on complex policies, we extend the Banerjee model to include this concern of differential quality by scaling the treatment effect by the county's GDP relative to the maximum, so that $TE_{adjusted,c} = TE_c \frac{GDP_c}{GDP_{maximum}}$.

The results from these simulations are displayed in Figure A.25. Mean t-statistics are (-0.001, 0.001, -0.001) for the Law enforcement for business, Care for left-behind children, and Tax classification policies respectively.

Experimental subject consent: If an experimental policy allows for subjects to opt-in (or opt-out), this may also induce selection in counties treated. We model this consideration in the simulation by only selecting treatment sites where the true treatment effect is greater than 0.[4]

The results from these simulations are displayed in Figure A.26. Mean t-statistics are (0.162, 0.052, 0.862) for the Law enforcement for business, Care for left-behind children, and Tax classification policies respectively.

[3] In practice, the expected value of the optimal experimental design and RCT may be equal for a given prior due to the discrete nature of the prior distribution. In these cases, we assign the 'indicator' variable for optimal RCT vs. deterministic design a value of 0.5 and take the t-statistic from the deterministic design.

[4] This places a strong assumption that counties know the true treatment of a given policy: introducing noise would weaken selection effects.

E Accounting for positive selection of experimentation sites

We argue that those political distortions indeed constitute a substantial part of the deviation from representative experimentation. To quantify the exact magnitude of deviation caused by those political concerns, we constructed a policy by prefecture dataset pooling all those features we explored in the previous sections, including political patronage, career incentive, and political unrest (from Section 4.4). For the baseline, we estimate the following econometric model using policy-prefecture level data:

$$y_{cp} = \alpha \cdot lngdppc_{cp} + Distortions_{cp}^{'}\beta + \gamma_p + \epsilon_{cp}. \tag{8}$$

Appendix Table A.15 shows the marginal effect of Log GDP per capita on the probability of being chosen as an experiment site. Positive selection bias is observed across columns. In columns 2 and 4, when those political distortions are controlled, the regression coefficients reduces to only half the amount without controls.

To answer this question from another direction, we ask ourselves how much deviation political distortions actually brings us. We begin with estimating a similar model as Equation 8, but without the explicit GDP per capita term. We then do a back-of-the-envelope calculation computing the prior probabilities (the propensity scores) of prefectural units receiving chances of experimentation given their level of distortion.

Appendix Figure A.27, Panel B shows the distribution of *t* statistics of the representative test, as described in Section 4.1, when we assert a non-stochastic version of treatment assignment mechanism. In this setting, those prefectural units with the top *k* propensity score get chosen as experimentation spots, where *k* corresponds to the number of sites chosen for each policy at status quo. Compared with our baseline specification shown in Appendix Figure A.27, Panel A, we observe positive selection bias of even greater magnitude. This is consistent with the strict nature of the non-stochastic assignment of policy experimentation.

For a milder version, we plot the distribution of *t* statistics of the representative test, when we assign experimentation sites in a stochastic fashion, according to their fitted propensity scores within each policy. For simplicity, we assume the sampling procedure is i.i.d., and the number of experimentation sites remains the same as that chosen at status quo. We conduct 1,000 simulations and plotted the pooled results in Appendix Figure A.27, Panel C. This specification is most similar, in general ideas, to the regression presented in Table A.15, confirming the idea that all distortion factors we identified explain almost half of the selection bias of policy experimentation.

F Measuring similarity among policy documents for local implementation of the experimentation

Following Bertrand et al. (2020), we load the corpus, split the text to break it into discrete tokens, count the number of times each token occurs in each document, and the number of documents in which each token occurs. We use *jieba*, a standard library widely used in Chinese NLP tasks. See github.com/fxsjy/jieba for details.

We use a standard Chinese stop-word library to clean up the tokens that are too frequent to be informative, and then encode each count of token i in document j into a feature weight w_{ij} with a common form of TF-IDF weighting.

$$w_{ij} = c_{ij} \ln \left(\frac{N}{n_i} \right),$$

where n_i is the number of documents containing at least one occurrence of token i, and N is the total number of documents in the corpus. We then stack these weights into a large, sparse, feature-document matrix M and apply a truncated singular value decomposition (SVD) to compute a rank D approximation of M:

$$\begin{gathered} U_D \Sigma_D V_D' = \arg\min \|M - M_d\|_F^2 \\ \text{s.t.} \quad \text{rank}(M_d) = D \end{gathered}$$

We discard U_D and take the singular value-scaled matrix $\Sigma_D V_D'$ as our set of Latent Semantic Analysis (LSA) document vectors. The word latent in LSA refers to the idea that compressing the full feature-document matrix to a lower-dimensional approximation often squeezes synonyms and other co-occurring words into the same singular vectors, improving the quality of the document model. The choice of D is often a matter of cross-validation. In our baseline model, we set $D = 3$.

G Additional figures and tables

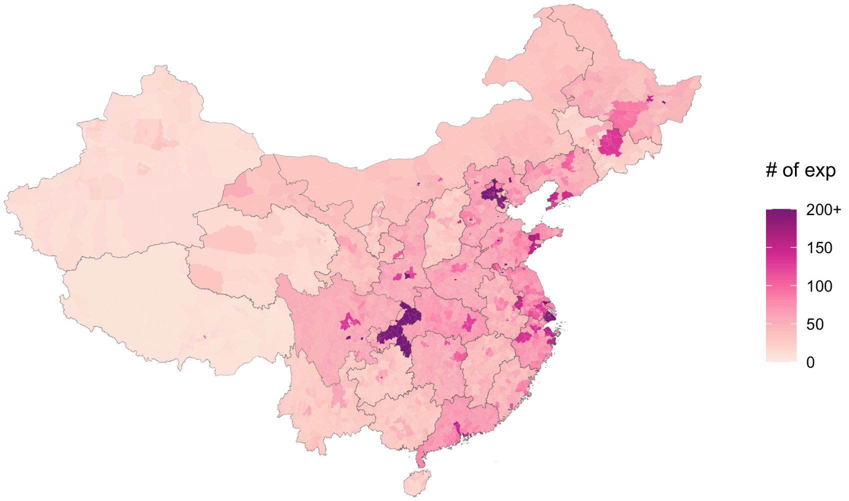
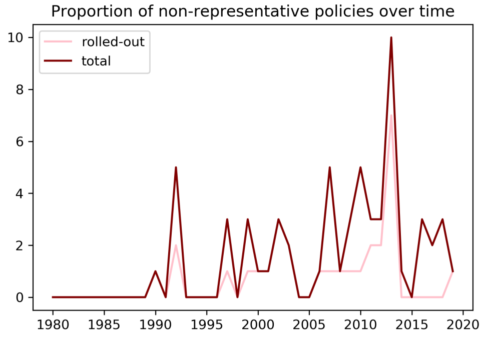
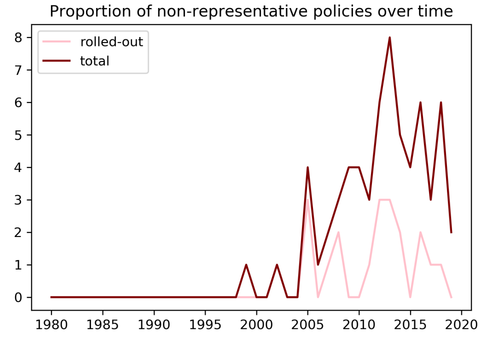


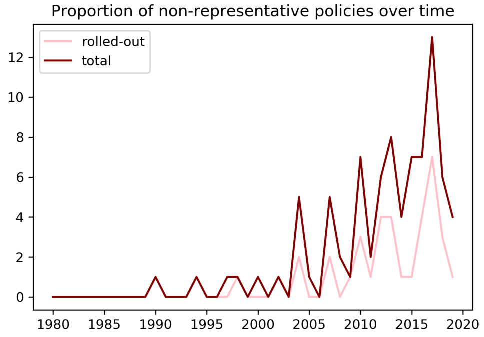
Figure A.1: This county-level map plots the spatial distribution of policy experimentation in China. A county is assigned a policy experiment if either itself or its corresponding prefecture/province serves as an experimentation site for that policy.



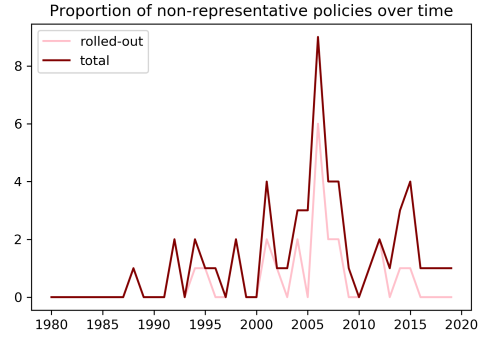
Education



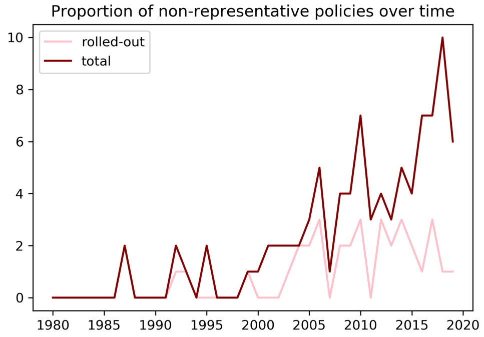
Agriculture



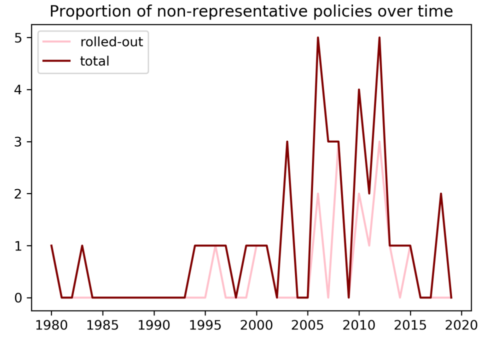
Market supervision



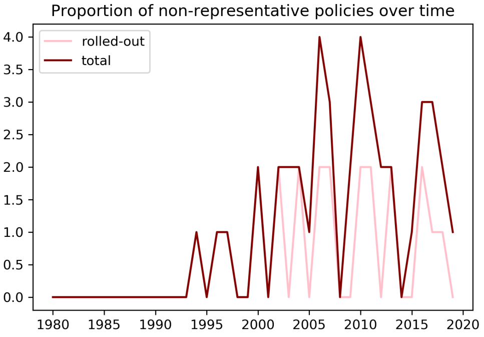
Finance & economics



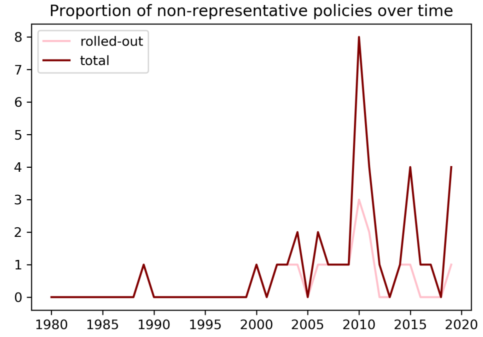
Natural resources & environment



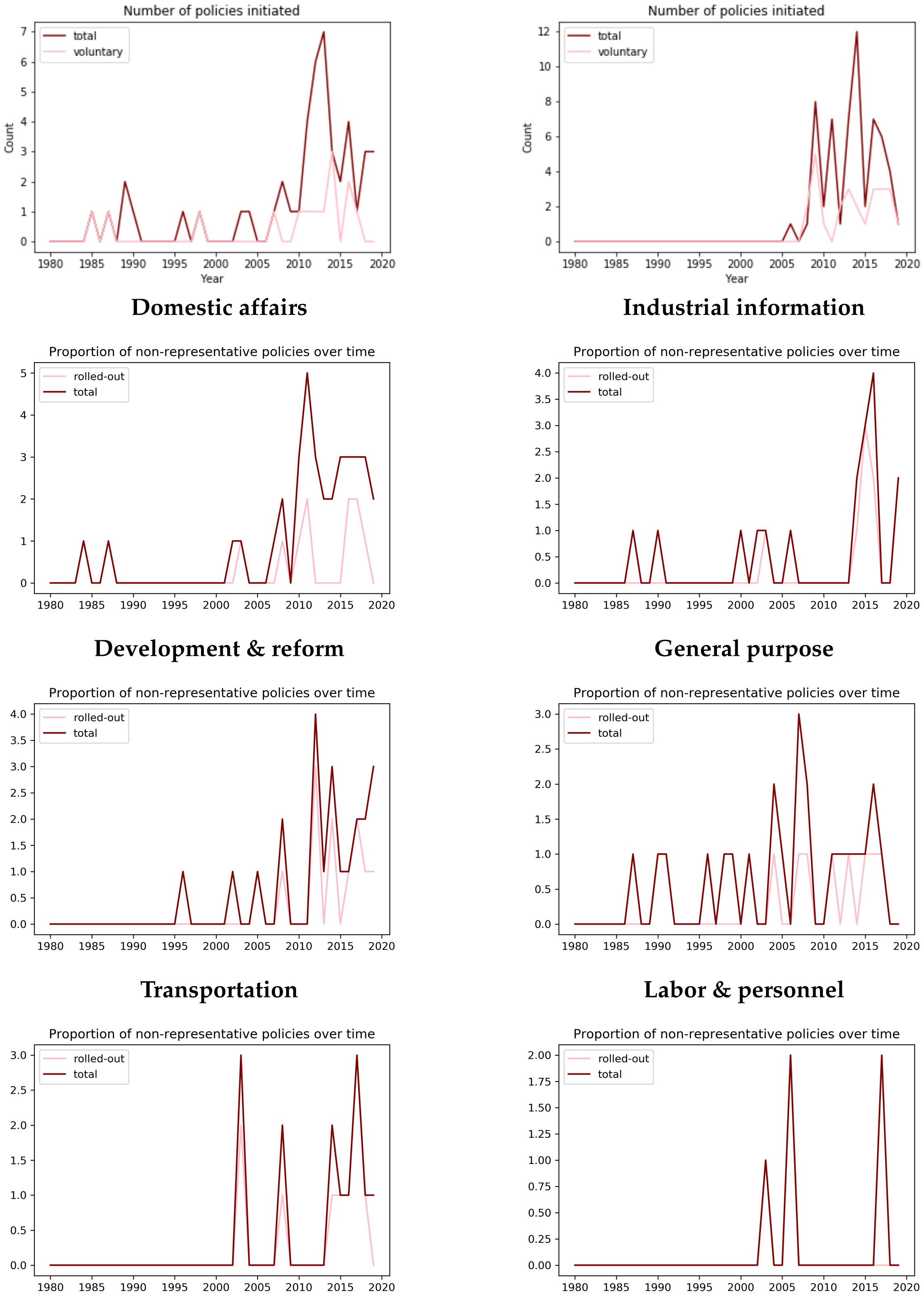
Business & commerce



Government finance & taxation



Population & health



Domestic affairs

Industrial information

Development & reform

General purpose

Transportation

Labor & personnel

Judiciary & supervision

Media

Figure A.2: Count of policy experimentation, by ministerial function.

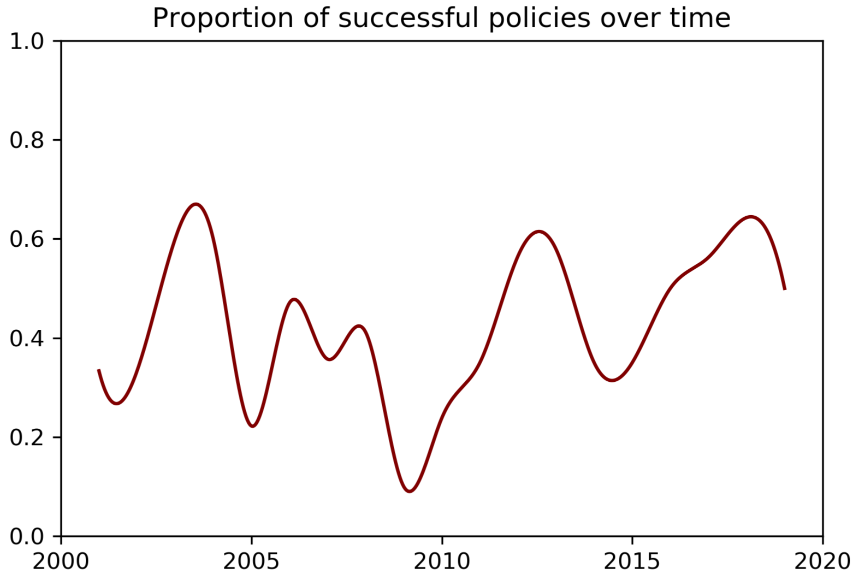


Figure A.3: This figure plots the ratio of successful policy experiments in each year. A policy experiment is defined as a “success" if it eventually rolled out to the entire nation.

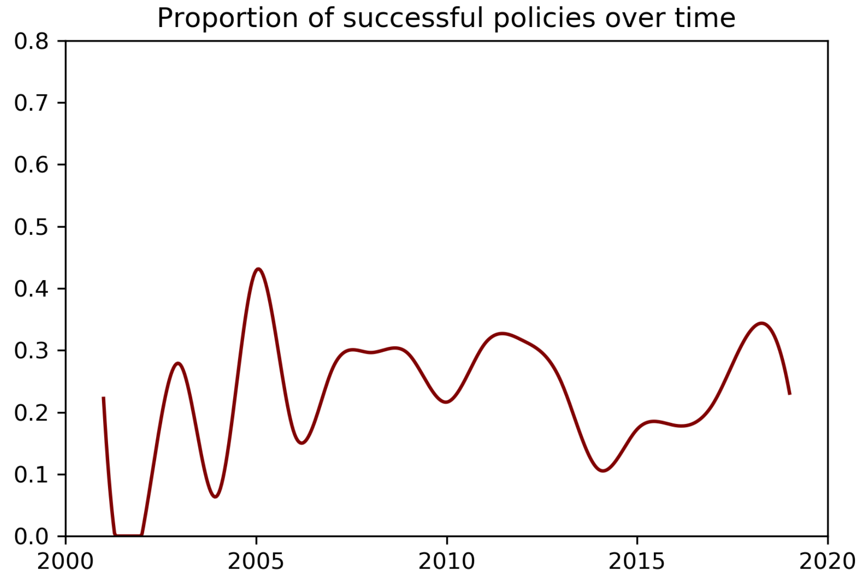


Figure A.4: Time trend of successful policy shares, 2000-2020. A policy is defined "success" if it is adopted by 2/3 of the provinces during the experimentation.

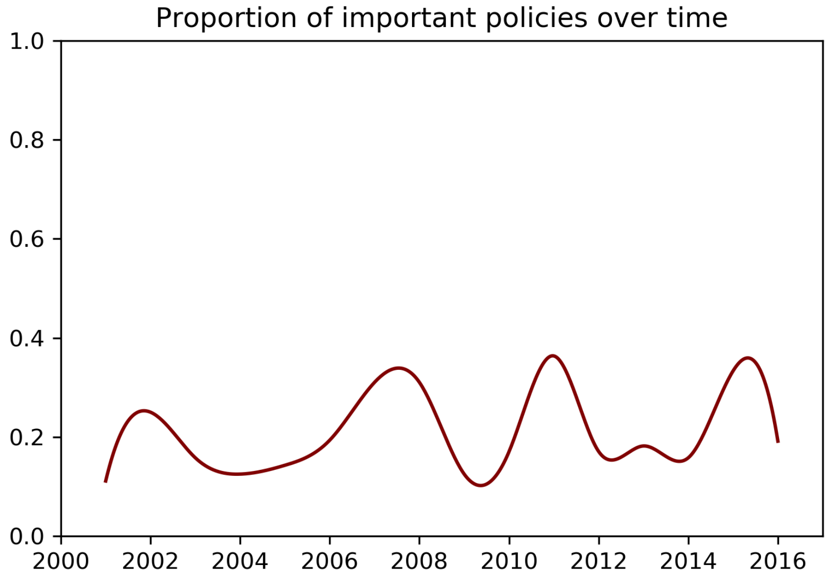


Figure A.5: Time trend of important policy shares, 2000-2017. A policy is labeled important if it is mentioned in the Five Year Plans. Those plans are both retrospective and introspective. We dropped some of the most recent years, observing that the most recent Five Year Plan is issued in 2016.

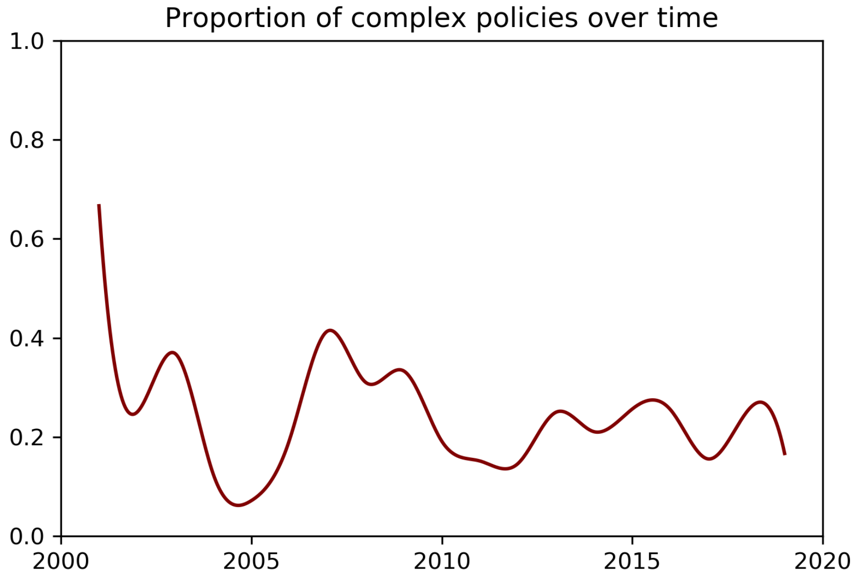


Figure A.6: This figure plots the share of policy experiments in each year that requires multi-department cooperation.

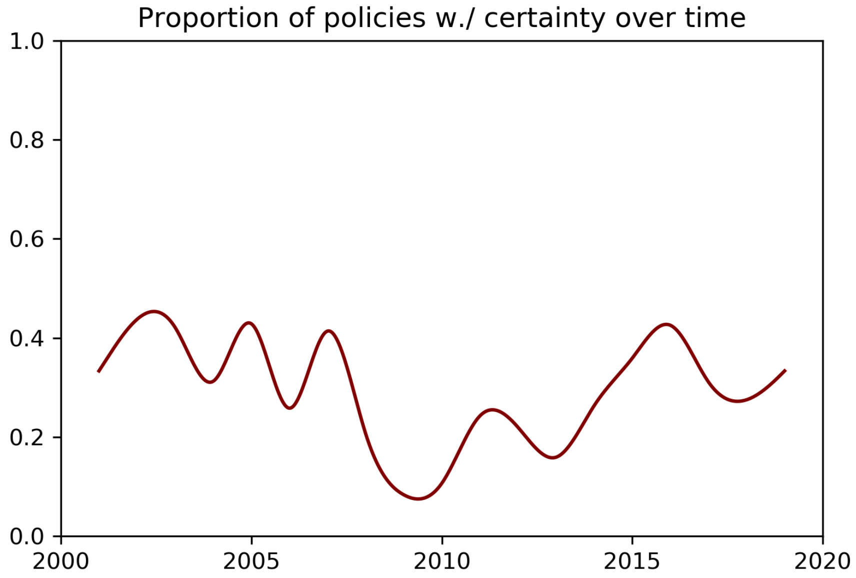


Figure A.7: This figure plots the share of policy experiments in each year that has detailed timelines of roll-out delineated in the first experimentation document.

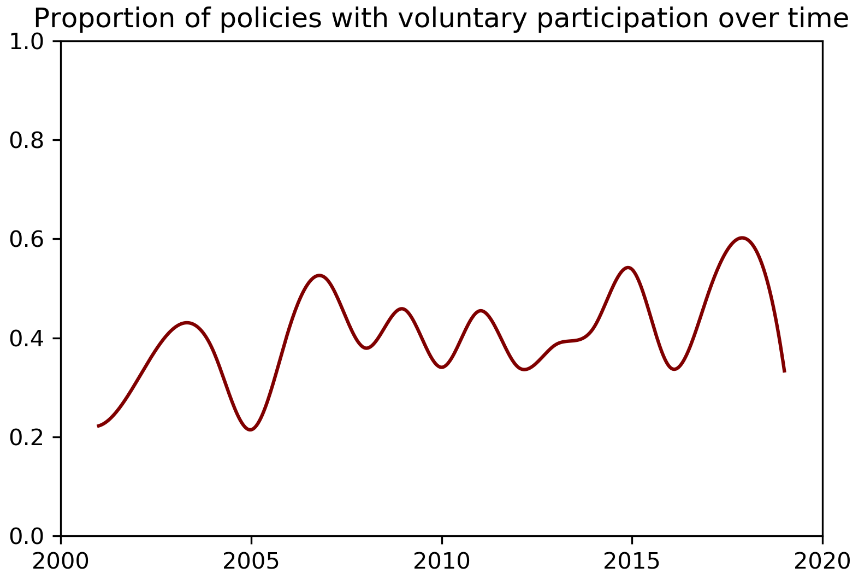


Figure A.8: This figure plots the share of policy experiments in each year that has a voluntary sign-up process for experimentation sites.

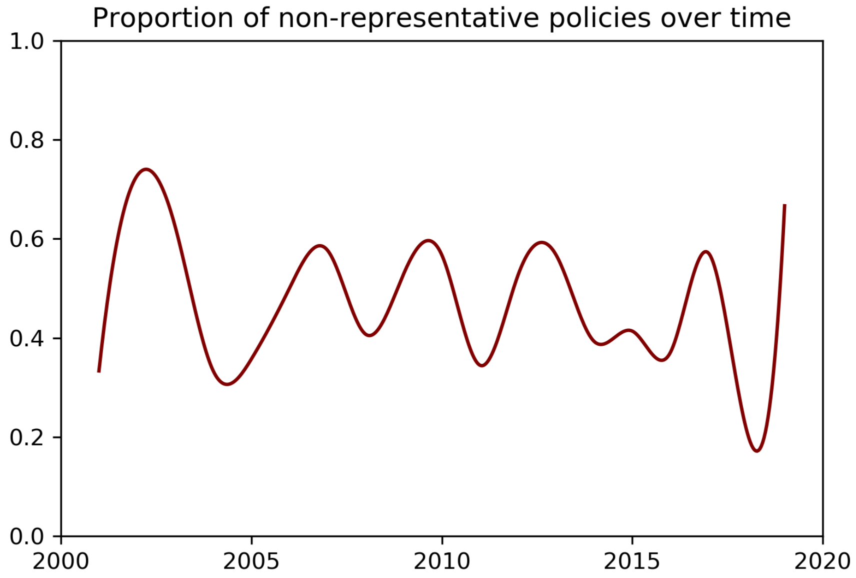
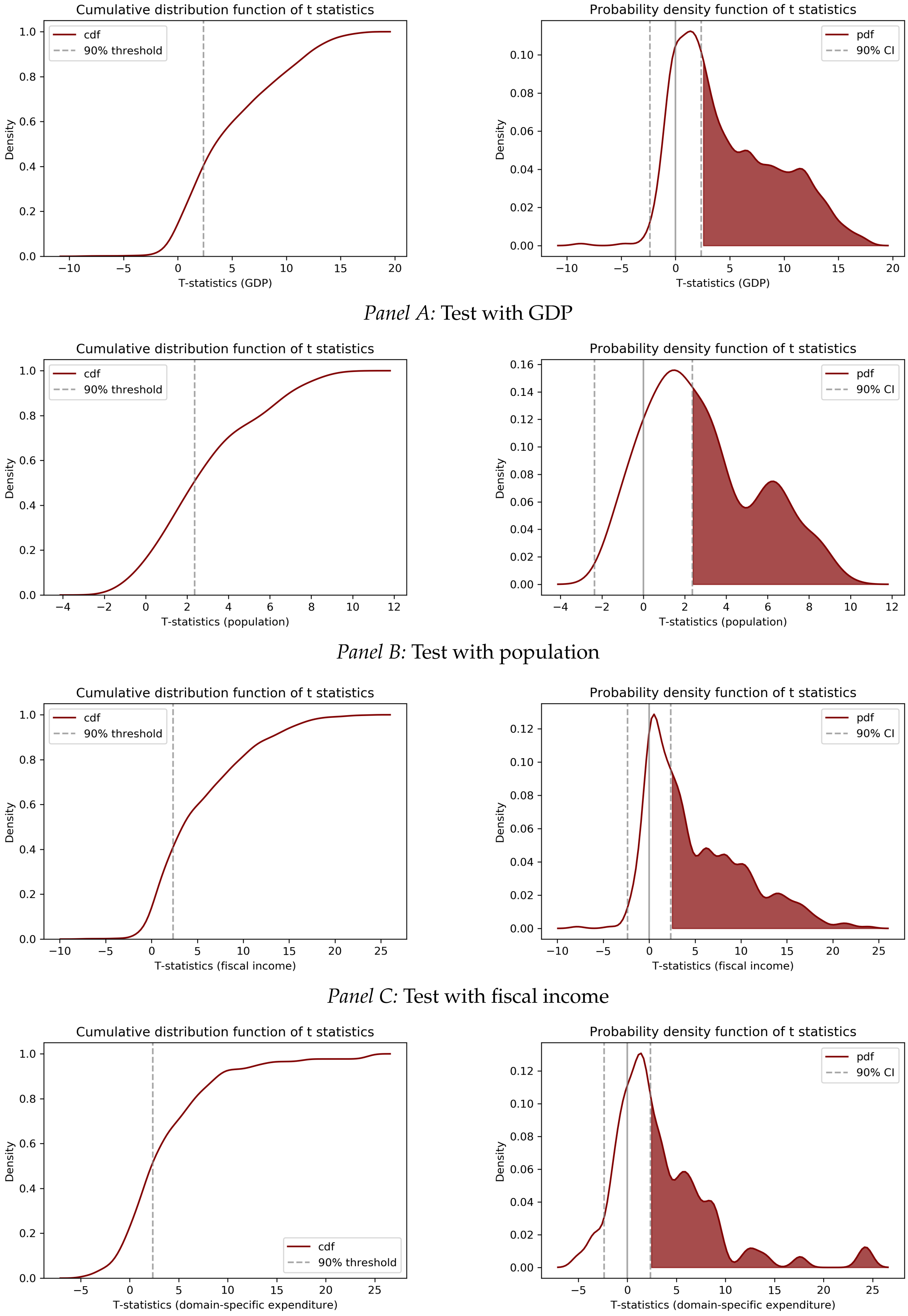


Figure A.9: This figure plots the share of non-representative policy experiments in each year. The notion non-representativeness is defined in Section 4.1.



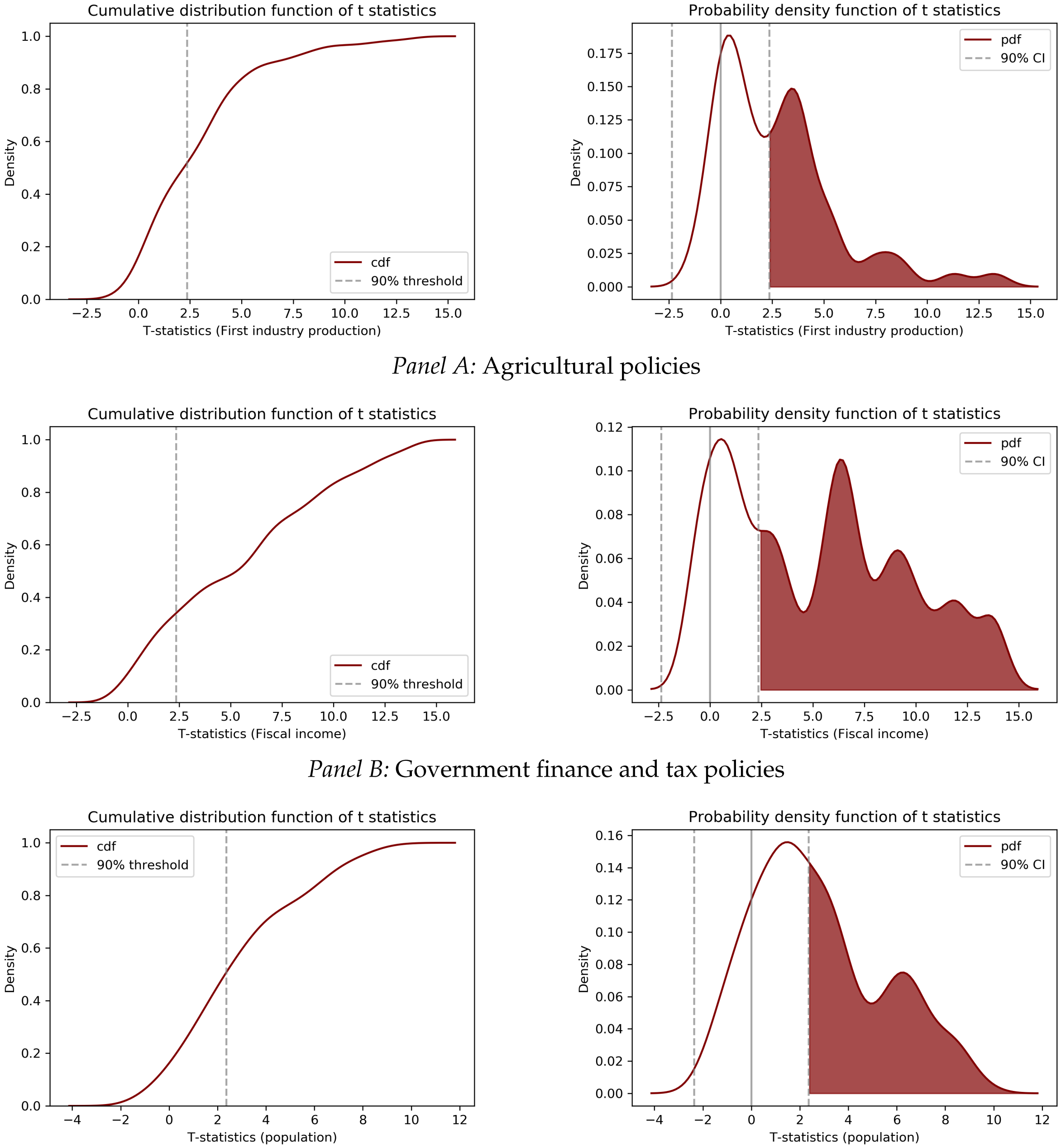
Panel A: Test with GDP

Panel B: Test with population

Panel C: Test with fiscal income

Panel D: Test with domain-specific fiscal expenditure

Figure A.10: This figure presents representativeness tests using alternative measures. In addition to the GDP per capita test as addressed in Figure 3, we conduct the test using total GDP in Panel A, population in Panel B, fiscal income in Panel C, and domain-specific fiscal expenditure in Panel D.

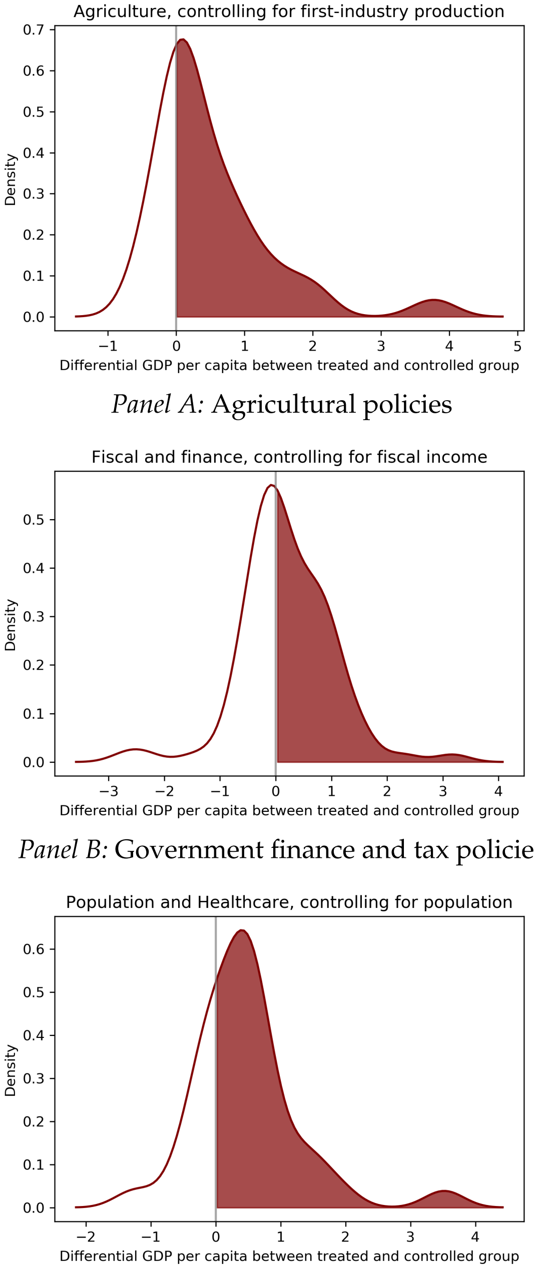


Panel A: Agricultural policies

Panel B: Government finance and tax policies

Panel C: Population and health policies

Figure A.11: This figure presents domain-specific representativeness tests. In Panel A, we focus on the subset of policies issued by the Ministry of Agriculture. We test for balance in pre-experimentation gross first-industry product. In Panel B, we focus on the subset of policies issued by the Ministry of Finance, and test for balance in pre-experimentation formal fiscal income. In Panel C, we explore the policies issued by the Ministry of Health and the National Population and Family Planning Commission and directly tested the population size against each other.

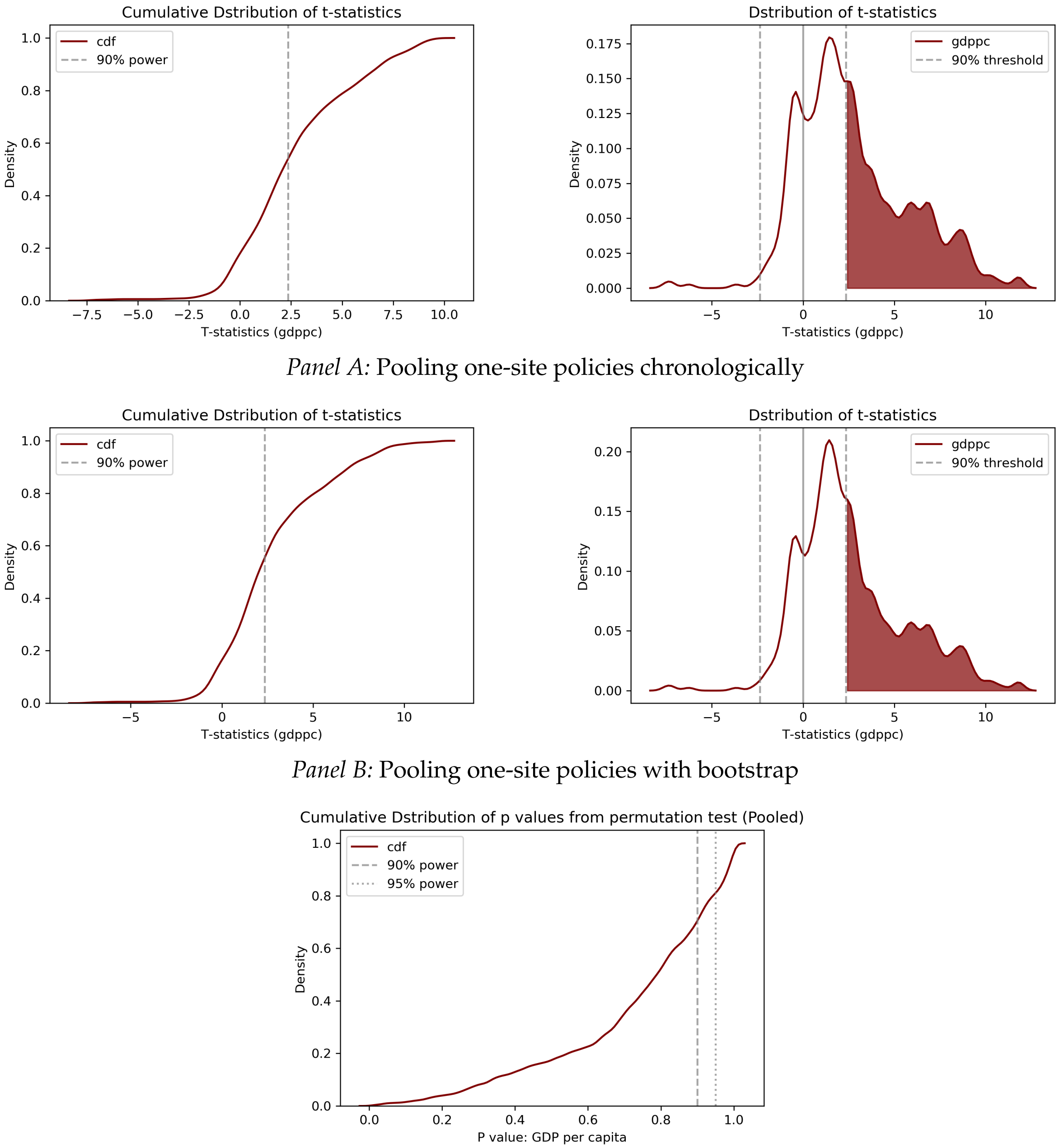


Panel A: Agricultural policies

Panel B: Government finance and tax policies

Panel C: Population and health policies

Figure A.12: This figure presents an alternative test specification where we address an assumption rationalizing positive selection by focusing on different units of analysis. We plot the distribution of regression coefficients of the treatment dummy, in a specification with controls that are believed to be highly correlated with the unit of analysis. In Panel A, we focus on the subset of policies issued by the Ministry of Agriculture, and controlled for the gross production of agricultural industry. We test for balance in pre-experimentation GDP per capita in a regression form. In Panel B, we focus on the subset of policies issued by the Ministry of Finance, and test for balance in GDP per capita, controlling for fiscal income. In Panel C, we explore the policies issued by the Ministry of Health and the National Population and Family Planning Commission and tested for GDP per capita, controlling for population.



Panel A: Pooling one-site policies chronologically

Panel B: Pooling one-site policies with bootstrap

Panel C: Incorporating one-site policies with a standard permutation test

Figure A.13: This figure presents robustness checks for the baseline representativeness test. Panels A and B are the same representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. In Panel A, one site policies are pooled chronologically by clusters n=5; in Panel B, one-site policies are pooled by bootstrapping with replacement for 166 times, without specifying the existence of any certain policy, and concatenated to the multi-site sample. Panel C shows the cumulative distribution of p values from permutation tests in representativeness tests. Each realized student-t statistic is compared with 5,000 permuted t values to calculate the p statistic. In small samples, permutation tests are more conservative than standard t tests.

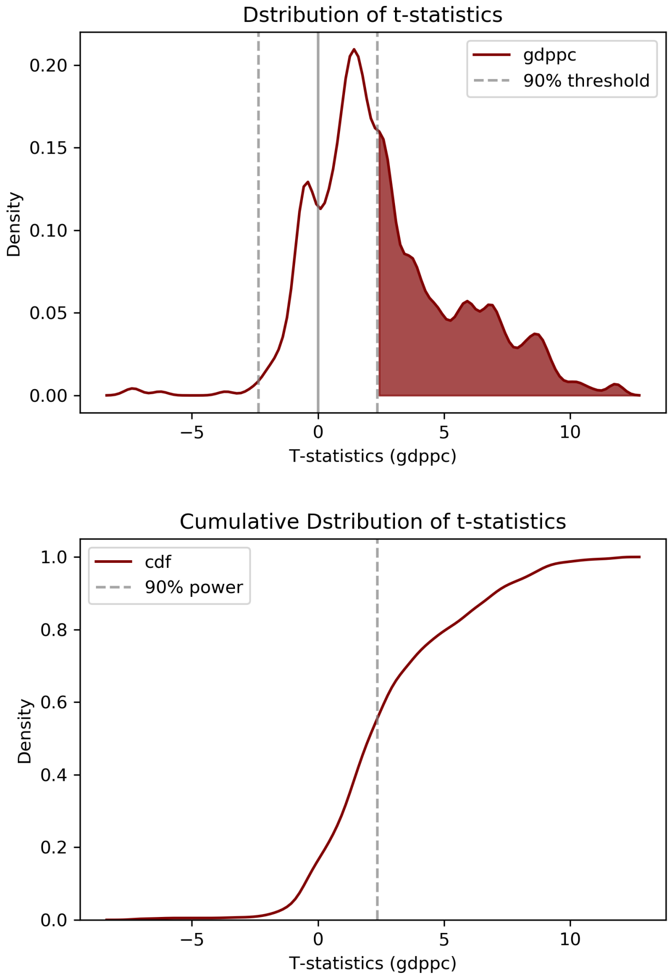


Figure A.14: This figure presents the representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. Only policies in early rounds are considered.

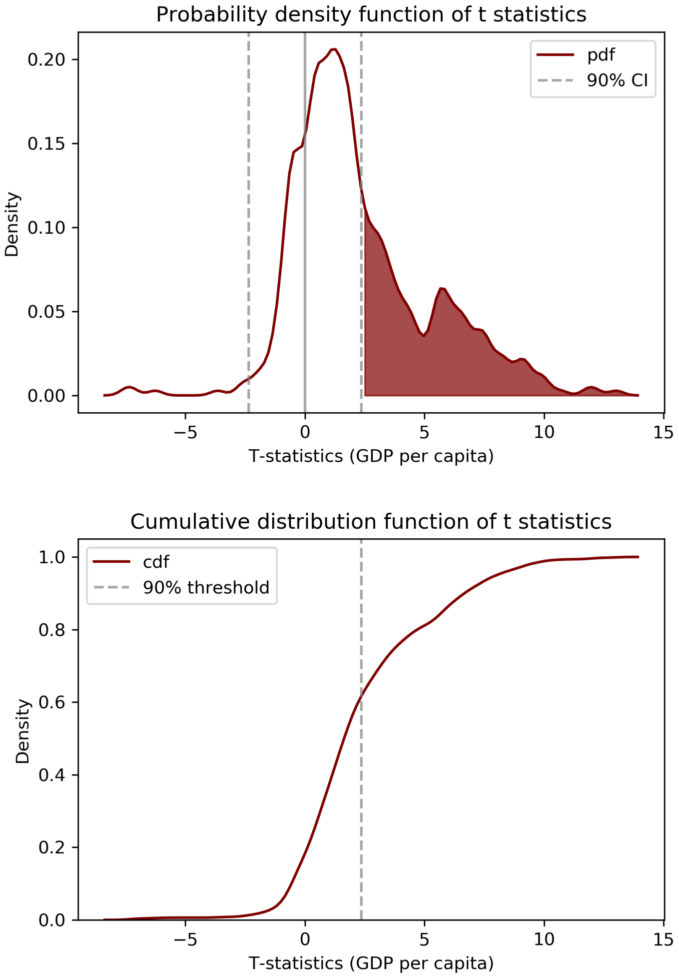


Figure A.15: This figure presents the representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. Municipalities are excluded from both treatment sample and control group

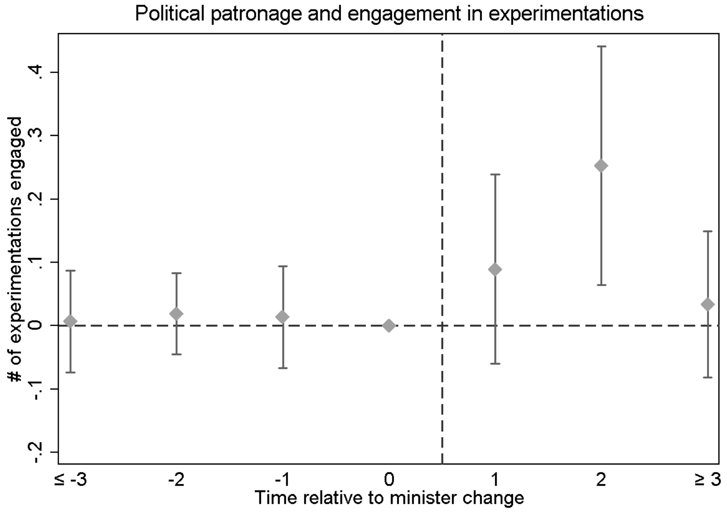


Figure A.16: This figure plots the event study estimates on a province's probability of being selected as an experimentation site after it becomes connected to a ministry due to political turnovers at the ministerial level.

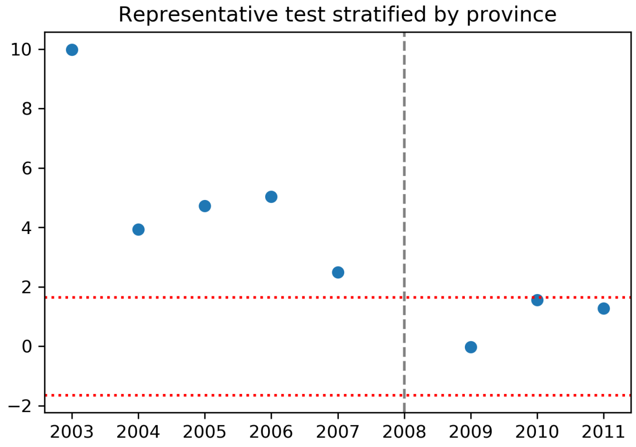


Figure A.17: Local fiscal reform — representativeness test. We conduct stratified Fisher randomization tests with student-t statistics and provincial strata. Within each province, we view counties that engage in the experimentation for the first time as units of the treatment group, the rest as control. Provincial level t-stats are weighted and standard errors are estimated based on Miratrix, Sekhon, and Yu (2013). The red horizontal lines indicate the asymptotic 90% confidence intervals within which representative assignment of experimentation sites cannot be rejected.

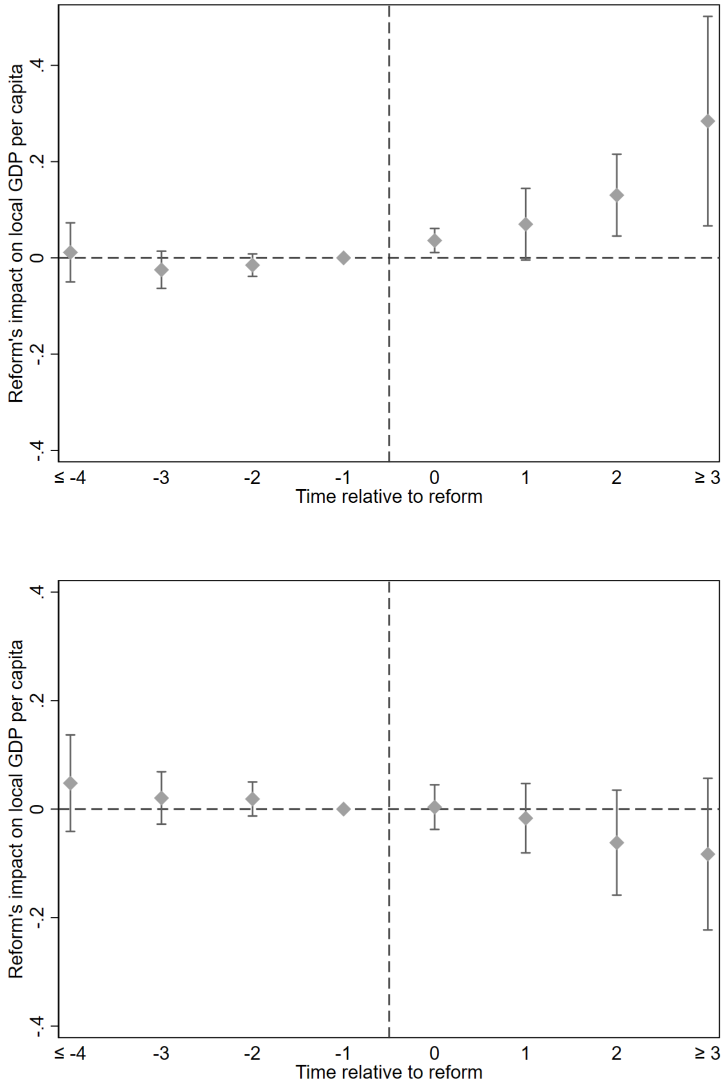


Figure A.18: Local fiscal reform - treatment effect on local GDP per capita. Upper Panel plots the counties reformed before 2007 and Lower Panel plots those reformed after 2007.

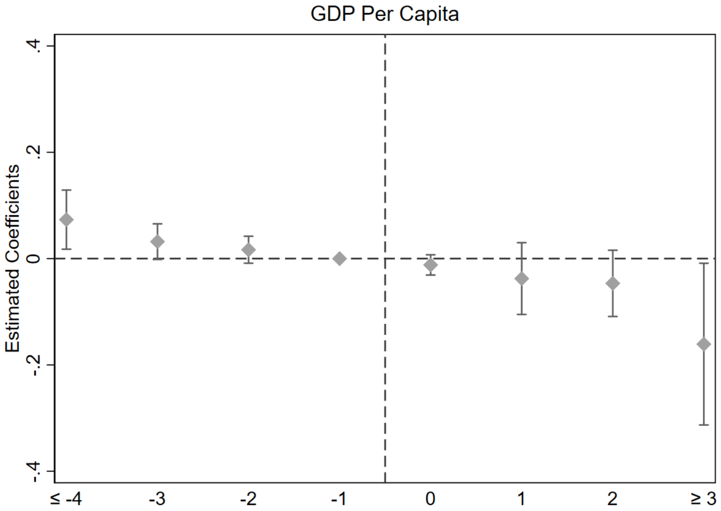


Figure A.19: Local fiscal reform - poor counties experimented before 2008. The pattern demonstrated here is similar to that in Figure A.18, Panel B.

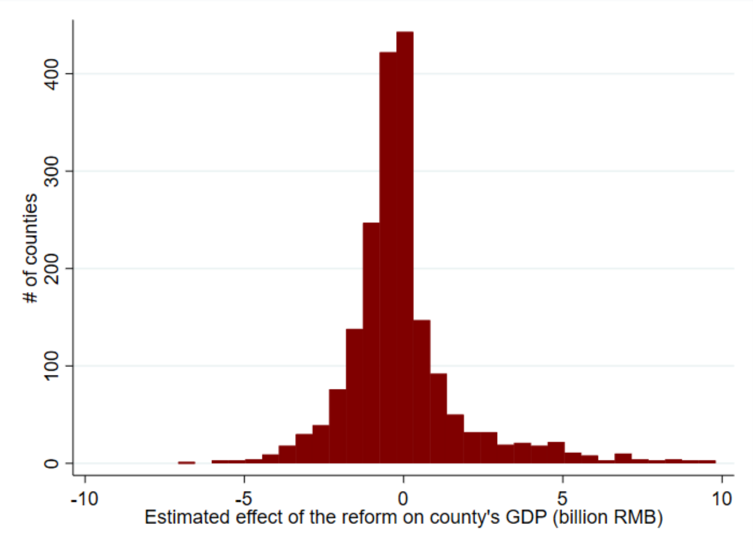


Figure A.20: Simulated treatment effects across country. We extrapolate the estimated treatment effect to all counties nationwide and obtain a distribution of reform effect on county's GDP.

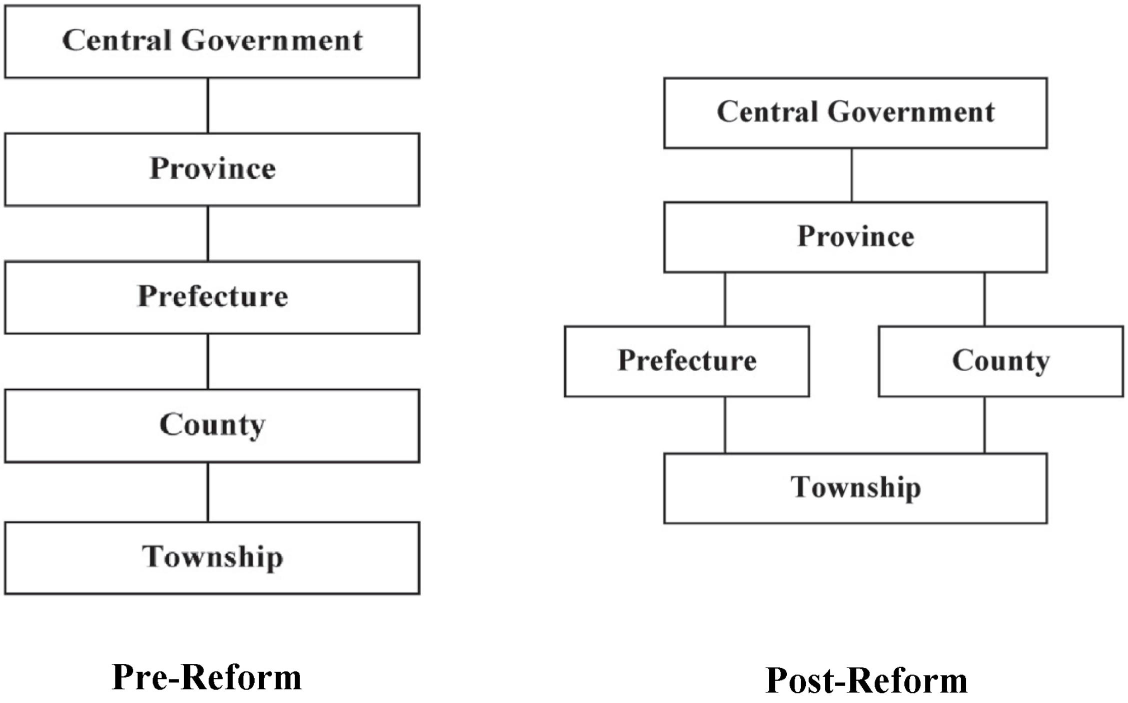


Figure A.21: Reproduced from Li, Lu, and Wang (2016). Illustration of local fiscal reform. After the reform, the provincial government could directly manage some of its counties, bypassing the prefectural cities, which grants county governments with more fiscal autonomy.

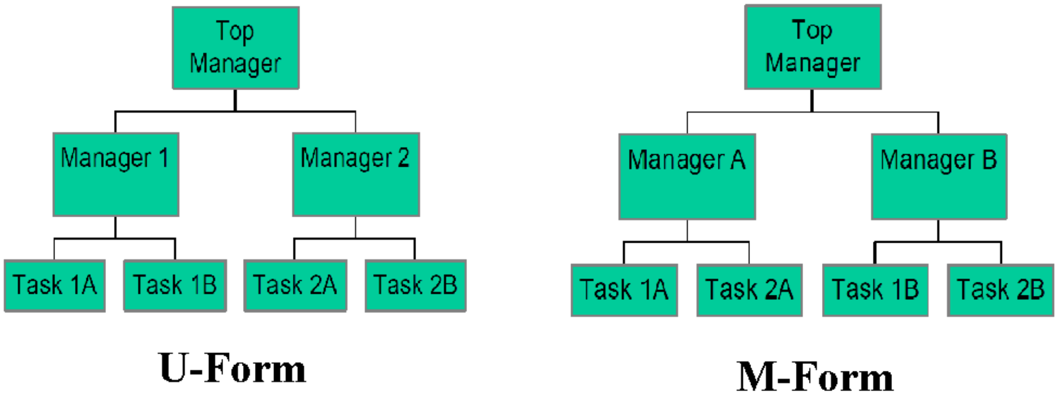


Figure A.22: Reproduced from Qian, Roland, and Xu (2006). Illustration of a shift from M form to U form. The top manager (ministers at central government branches in our case) have more administrative and personnel authority on its branches.

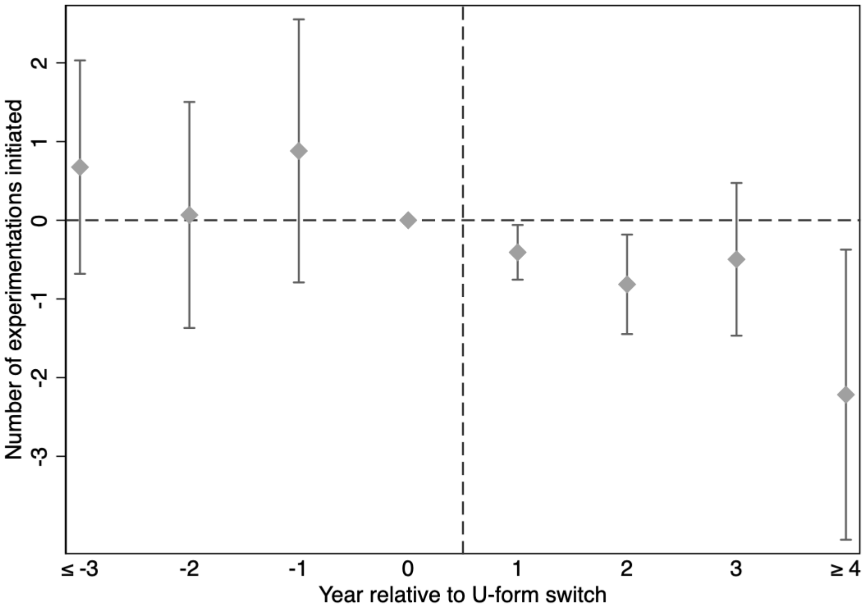


Figure A.23: Count of policy experimentations initiated after transitioning into U-form. X-axis indicates the time relative to the reform. The point estimates and confidence intervals are computed from a standard event study design controlling for ministry fixed effect and calendar year fixed effect. Standard errors are clustered at the ministry level.

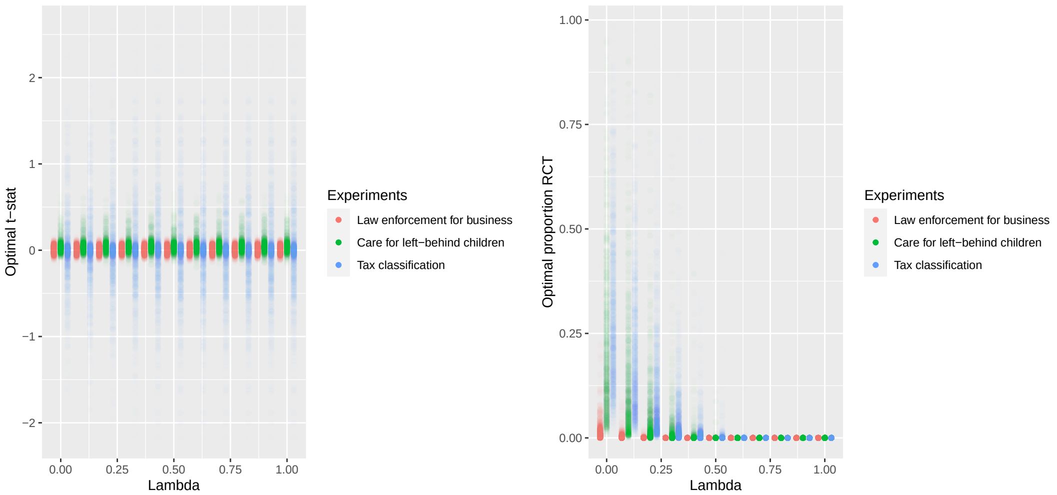


Figure A.24: This plot shows optimal t-statistics (left) and RCT vs. deterministic experimental policy (right) for simulations calibrated using three different policy experiments conducted in China following the model in Banerjee et al. 2020. Lambda ranges from 1 (full weight on decision maker's utility) to 0 (full weight on most adversarial prior). Mean t-statistics are (0.006, 0.051, -0.006) for the Law enforcement for business, Care for left-behind children, and Tax classification policies respectively.

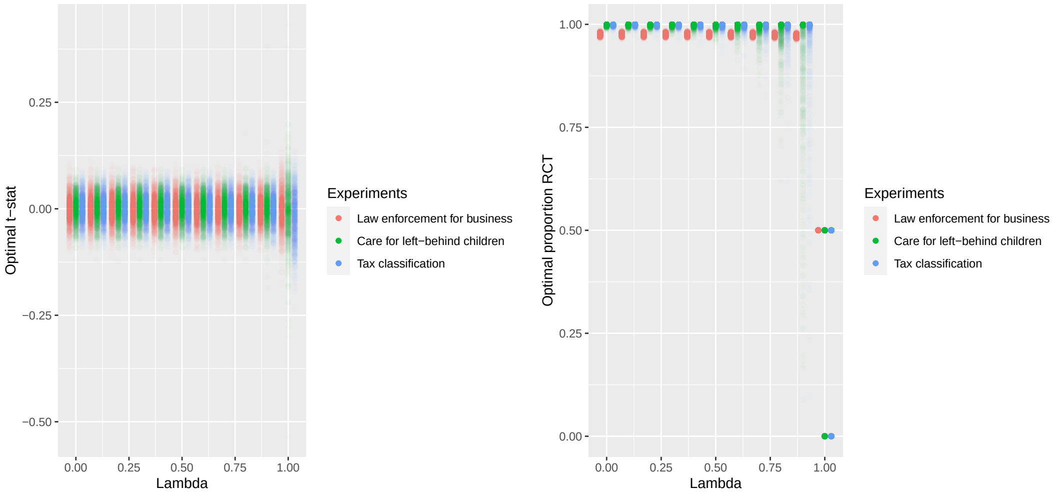


Figure A.25: This plot shows optimal t-statistics (left) and RCT vs. deterministic experimental policy (right) for simulations calibrated using three different policy experiments conducted in China following the model in Banerjee et al. 2020 with differential quality of information. Lambda ranges from 1 (full weight on decision maker's utility) to 0 (full weight on most adversarial prior). Mean t-statistics are (-0.001, 0.001, -0.001) for the Law enforcement for business, Care for left-behind children, and Tax classification policies respectively.

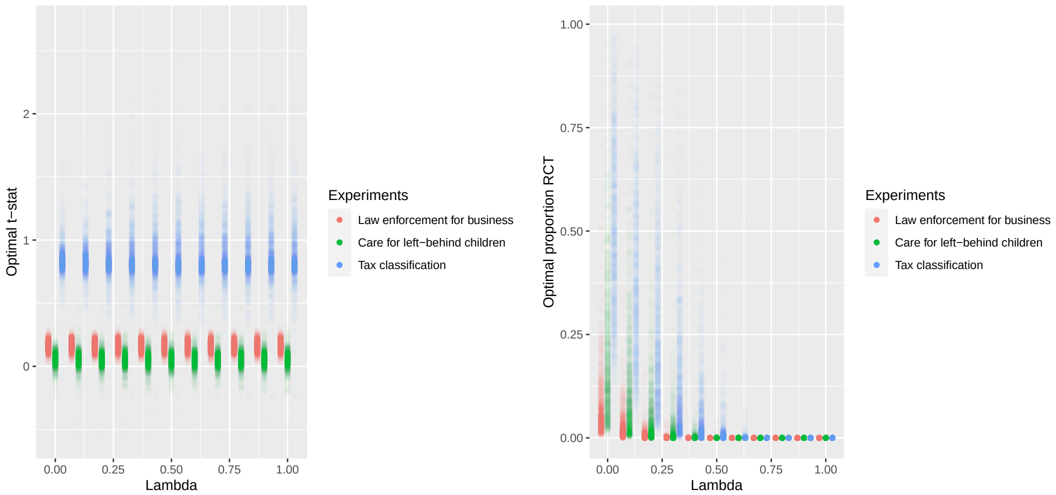


Figure A.26: This plot shows optimal t-statistics (left) and RCT vs. deterministic experimental policy (right) for simulations calibrated using three different policy experiments conducted in China following the model in Banerjee et al. 2020 with subject consent. Lambda ranges from 1 (full weight on decision maker's utility) to 0 (full weight on most adversarial prior). Mean t-statistics are (0.162, 0.052, 0.862) for the Law enforcement for business, Care for left-behind children, and Tax classification policies respectively.

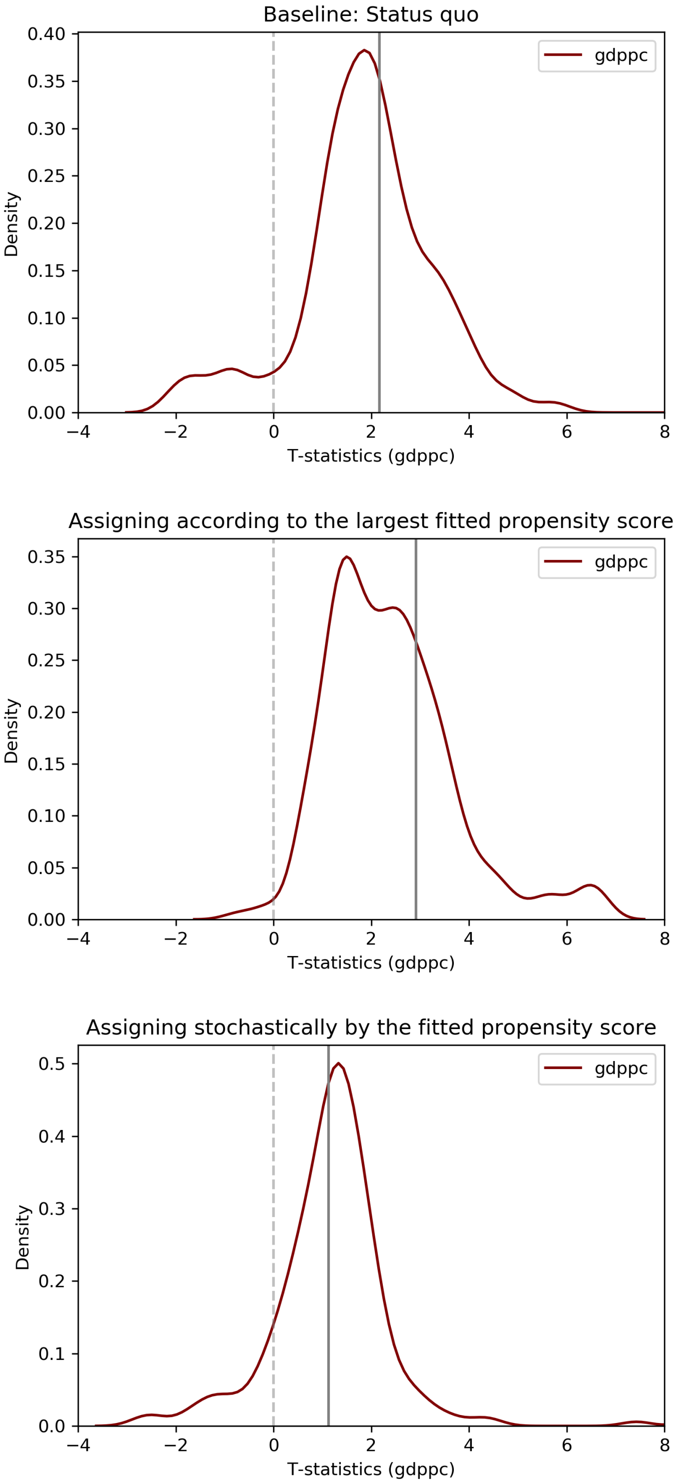


Figure A.27: Representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. To make sure we're making reasonable comparisons, we adjust the baseline Panel A so that it only includes the experimentations targeting prefectural cities, with 4 municipal cities excluded from observation. Panel B shows the simulated results when sites for experimentation are assigned to the prefectural units with the largest fitted propensity score. Panal C imposes a milder assumption where we assert a stochasitic version. Extreme values are trimmed. The grey vertical line indicates the mean of t-stats.

Table A.1: Comprehensive checks for *PKUlaw* dataset

Ministry	Official #	*PKULaw* #	Coverage
	(1)	(2)	(3)
State Council	1066	1082	92.8%
Environment	111	99	91.0%
Fiscal	192	371	88.5%
Natural Resources	181	230	86.7%
Education	854	1053	78.0%

Note: In columns 1 and 2, we respectively report the number of all central policy documents issued by the ministry available on the website. Column 3 we report the ratio of experimentation-related policy documents issued by the central government that is found with its exact title in the *PKULaw* database. We then manually iterate through them. The numbers reported are very conservative. Fixing encodings of annotations and dropping secondary documents irrelevant to experimentation will give us a larger ratio, but for consistency we do not report the calibrated numbers. In most cases, *PKULaw* collects even more documents than the official websites. One complication is that some of the ministries only publicized their policies in very recent years (e.g., Fiscal and Tax; Natural Resources). To address this issue, we confine the numbers of policies that are compared against to the same time frame.

Table A.2: Changes in positive experimentation sites selection over time

	Year		
	coef.	s.e.	coef / mean
	(1)	(2)	(3)
Panel A: Full sample			
OLS	-0.067	0.024	-0.025
Ministry FE	-0.074	0.031	-0.028
Panel B: By ministry			
Industry & information technology	-0.449	0.224	-0.112
Transportation	-0.114	0.113	-0.097
Agriculture	-0.174	0.083	-0.096
Labor & personnel	-0.156	0.094	-0.07
Tax & fiscal policy	-0.161	0.109	-0.058
Law	-0.18	0.182	-0.051
Development & reform	-0.091	0.125	-0.049
Commerce & trade	-0.121	0.092	-0.031
Education	-0.07	0.061	-0.029
Population & health	-0.046	0.086	-0.022
Finance	-0.111	0.15	-0.021
Resource, energy & environment	0.006	0.045	0.003
Market supervision	0.068	0.052	0.021
Domestic affairs	0.157	0.078	0.097
State ministry	0.262	0.266	0.102

Notes: The tables shows results regressing t-stats on calendar year. We report the coefficients in column 1, robust standard errors in column 2, and the coefficients relative to within ministry mean in column 3.

Table A.3: Engagement in experimentation and local politicians' promotion

	Promotion			
	(1)	(2)	(3)	(4)
Panel A: All politicians				
Participated in experimentation (all)	-0.025	-0.040		
	(0.053)	(0.057)		
Participated in experimentation (rolled-out)			0.087***	0.098***
			(0.031)	(0.034)
# of obs.	1139	1139	1139	1139
Mean of DV	0.369	0.369	0.369	0.369
Prefecture FE	No	Yes	No	Yes
Panel B: Politicians with above median career incentives				
Participated in experimentation (all)	-0.012	-0.034		
	(0.071)	(0.083)		
Participated in experimentation (rolled-out)			0.191***	0.166***
			(0.043)	(0.052)
# of obs.	586	586	586	586
Mean of DV	0.433	0.433	0.433	0.433
Prefecture FE	No	Yes	No	Yes

Note: Standard errors are clustered by prefectures in column 2 and 4. We explore whether strategic efforts and experimentation-engagement are correlated with politician's promotion. We group our observations to city level to match the career trajectory information.

Table A.4: Falsification test: experimentations and pre-period career incentive

	Engage in experimentation		
	(1)	(2)	(3)
Immediate predecessor's career incentive	-0.697	-0.724	-0.464
	(0.507)	(0.505)	(0.484)
# of obs.	5857	5857	5857
Mean of DV	1.028	1.028	1.028
Prefecture Controls	No	No	Yes
Politician Controls	No	Yes	Yes
Prefecture FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes

Note: Standard errors clustered at the prefecture level are reported in parentheses. This exercise is parallel to Table 2, Panel B, and here we conduct falsification test by substituting in-office city leader's career incentive with that of his immediate predecessor.

Table A.5: Political career incentives and engagement in experimentation

	Engaged in experimentation		
	(1)	(2)	(3)
Panel A: All experiments			
Career incentive	1.397*	1.405*	1.309*
	(0.796)	(0.824)	(0.780)
Panel B.1: Experiments initiated by M-form ministry			
Career incentive	1.541**	1.561**	1.467**
	(0.674)	(0.696)	(0.686)
Panel B.2: Experiments initiated by U-form ministry			
Career incentive	0.181	0.186	0.185
	(0.139)	(0.143)	(0.142)
Panel C.1: Experiments with top-down assignments			
Career incentive	0.721*	0.703*	0.664
	(0.400)	(0.418)	(0.410)
Panel C.2: Experiments with voluntary sign-ups			
Career incentive	0.676	0.702	0.642
	(0.517)	(0.534)	(0.528)
# of obs.	7630	7630	7630
Mean of DV	1.059	1.059	1.059
Prefecture controls	No	No	Yes
Politician controls	No	Yes	Yes
Prefecture FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes

Note: The standard errors clustered by prefectures are reported below the estimates. Control variables at the politician level include the educational level and previous central-government positions. Control variables at the prefecture level include GDP per capita, fiscal income, and fiscal expenditure, all in logarithms. We also controlled for the career incentive of the previous city leader to address the concern where the engagement is just a continuation of previous progress. The construction of career incentive index is introduced in Appendix Section B.1.

In panel A we report the estimated effect of career incentive intensity on all types of experimentation. In panel B we differentiate between experiments issued by a M-form ministry, where city leaders have direct control on the logistics of the policy; and experiments initiated by a U-form ministry where the central government takes direct orders on local branches. In Panel C, we investigate experiments with top-down assignments and voluntary sign-ups, respectively.

Table A.6: Political patronage and engagement in experimentation

	Engaged in experimentation		
	(1)	(2)	(3)
Panel A: All experiments			
Connected to minister	0.053**	0.037**	0.037**
	(0.020)	(0.016)	(0.016)
Panel B.1: Experiments initiated by M-form ministry			
Connected to minister	0.063***	0.025	0.025
	(0.021)	(0.017)	(0.017)
Panel B.2: Experiments initiated by U-form ministry			
Connected to minister	-0.001	-0.003	-0.004
	(0.048)	(0.047)	(0.047)
Panel C.1: Experiments with top-down assignments			
Connected to minister	0.054***	0.040**	0.040***
	(0.017)	(0.014)	(0.014)
Panel C.2: Experiments with voluntary sign-ups			
Connected to minister	0.020	0.011	0.011
	(0.018)	(0.017)	(0.017)
# of obs.	42884	42884	42884
Mean of DV	0.176	0.176	0.176
Controls	No	No	Yes
Year FE	No	Yes	Yes
Ministry by province FE	Yes	Yes	Yes

Note: The standard errors clustered at the province level are reported below the estimates. We control for ministry by province fixed effect in all regressions. Control variables include the provinces' value added of first and second industry, fiscal expenditure and income of local governments as control variables. The mean of dependent variable and count of observations of Panel A alone are reported.

In Panel A, we count all policy experiments, whereas in Panel B we distinguish between experiments initiated by M-form ministry and U-form ministry. Finally in Panel C, we investigate provincial engagement in experiments with voluntary sign-ups and top-down assignments, respectively.

Table A.7: Concerns for political stability and selection of experimentation sites

	Engaged in experimentation		
	(1)	(2)	(3)
Panel A: All experiments			
# of protests in previous year	-0.003***	-0.002**	-0.002***
	(0.001)	(0.001)	(0.001)
Panel B.1: Experiments initiated by M-form ministry			
# of protests in previous year	-0.003***	-0.002***	-0.002***
	(0.001)	(0.001)	(0.0003)
Panel B.2: Experiments initiated by U-form ministry			
# of protests in previous year	-0.001***	-0.001***	-0.001***
	(0.0003)	(0.0002)	(0.0001)
Panel C.1: Experiments with voluntary sign-ups			
# of protests in previous year	-0.004***	-0.003**	-0.004*
	(0.001)	(0.001)	(0.002)
Panel C.2: Experiments with top-down assignments			
# of protests in previous year	0.001	0.001	0.002
	(0.001)	(0.001)	(0.002)
# of obs.	1730	1730	940
Mean of DV	1.278	1.278	2.117
Pre-period controls	No	No	Yes
Year FE	No	Yes	Yes
Prefecture FE	Yes	Yes	Yes

Note: The standard errors clustered by prefectures are reported below the estimates. Prefectural fixed effects are controlled across columns. As discussed in Section 5, we witness significant selection bias of site selection in terms of GDP per capita. To account for this we controlled for GDP per capita, in logarithm, at prefecture level in the previous year in column 3.

Table A.8: Fiscal expenditure, incentive, and local economic growth

	GDP per capita growth			
	(1)	(2)	(3)	(4)
Fiscal expenditure	0.010***	0.048***		
	(0.001)	(0.005)		
Career incentive			0.002**	0.008***
			(0.001)	(0.002)
# of obs.	18,481	18,481	85,399	85,399
Mean of DV	0.173	0.173	0.126	0.126
County FE	No	Yes	No	Yes

Note: Standard errors for column 2 and 4 are clustered by county. We correlate the policy outcome, measured by GDP per capita growth, and effect measures such as total fiscal expenditure, and average career incentive. The former is observed at county level and the latter at prefectural city level. We map higher-level experimentations to all the localities within its jurisdiction.

Table A.9: Land revenue windfall and experimentation rollout - first stage

	Land revenue		
	(1)	(2)	(3)
Unsuitability× interest rate	3.353***	3.720***	3.661***
	(0.192)	(0.226)	(0.226)
# of obs.	16,967	16,967	16,967
Mean of DV	5.191	5.191	5.191
Controls	Yes	Yes	Yes
Ministry FE	No	No	Yes
Year FE	Yes	Yes	Yes
County FE	No	Yes	Yes

Note: The standard errors clustered at county level are reported below the estimates. Here, we show the first stage results for the two-stage-least-square regression in Table 5, panel A. The independent variable is the average land revenue collected, across the whole experimentation period, in logarithm level. We include politician level control variables including his or her age, education, past experience in the prefectural government, previous positions as Youth League party leaders, and hometown connection with the prefectural leaders.

Table A.10: Political rotation: falsification test

	National roll-out		
	(1)	(2)	(3)
Pre-exp rotation	-0.000	-0.000	-0.004
	(0.016)	(0.014)	(0.015)
Pre-exp rotation × change in career incentive	0.117	0.069	0.093
	(0.140)	(0.125)	(0.131)
# of obs.	2846	2842	2842
Mean of DV	0.261	0.261	0.261
Province FE	No	No	Yes
Ministry FE	No	Yes	Yes
Year FE	Yes	Yes	Yes

Note: Standard errors clustered at the province level are reported in parentheses. Here the specification is fully parallel to that in Table 5, Panel B. We consider political rotation in pre-experimentation period (the time window considered here is completely symmetric with respect to the start year of experimentation).

Table A.11: Representativeness of experimentation sites selection and policy's national roll-out

	National roll-out			
	Full sample	Full sample	Certain policies	Uncertain policies
	(1)	(2)	(3)	(4)
Non-representativeness	0.207***	0.228***	0.135	0.271***
	(0.058)	(0.063)	(0.127)	(0.068)
# of obs.	402	397	104	257
Mean of DV	0.568	0.568	0.764	0.477
Controls for hierarchical level	Yes	Yes	Yes	Yes
Controls for fiscal input	Yes	Yes	Yes	Yes
Ministry FE	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes

Note: The standard errors clustered at department level are reported below the estimates. Non-representativeness is an indicator of whether we can reject the null hypothesis that pre-experimentation GDP per capita is balanced between the experimented sites and the rest of the country.

Table A.12: Similarity with experimentation sites and effects of policy roll-out: Robustness check

	GDP per capita growth		
	(1)	(2)	(3)
Panel A: GDP per capita			
M-distance between local development	-0.006***	-0.007***	-0.006***
	(0.001)	(0.001)	(0.001)
Panel B: GDP per capita + Fiscal income			
M-distance between local development	-0.004***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)
Panel C: GDP per capita + Fiscal income + Population			
M-distance between local development	-0.004***	-0.004***	-0.003***
	(0.0003)	(0.001)	(0.001)
# of obs.	77,588	77,588	77,588
Mean of DV	0.0806	0.0806	0.0806
Policy FE	Yes	No	Yes
County FE	No	Yes	Yes

Note: Robust standard errors clustered at policy level are reported below the estimates. The idea of this table is illustrated in the notes of Table 6. The panel subtitles illustrates the different specifications of Mahalanobis distance we explored.

Table A.13: Ministries underwent vertical management reforms

Ministry	**Year**
China Securities Regulatory Commission	1998
People's Bank of China	1999
Ministry of State Security	2001
National Medical Products Administration	2001
Ministry of Natural Resources	2004
National Bureau of Statistics (Survey Team)	2004
State Administration for Coal Mine Safety	2005
State Post Bureau	2005
Ministry of Environmental Protection	2016

Table A.14: Predicting politicians' career incentives

	Promotion			
	OLS	OLS	Probit	Probit
	(1)	(2)	(3)	(4)
Start age	-0.019***	-0.013***	-0.051***	-0.037***
	(0.003)	(0.003)	(0.007)	(0.007)
hierarchical level	-2.201***	-2.148***	-6.417***	-6.355***
	(0.346)	(0.345)	(1.168)	(1.178)
Start age × hierarchical level	0.042***	0.040***	0.122***	0.118***
	(0.007)	(0.007)	(0.023)	(0.023)
Controls	No	Yes	No	Yes
# of obs.	2,337	2,337	2,337	2,337

Note: The robust standard errors are reported below the estimates. Control variables include the educational background of the city leader, and previous work experience in the central government. We do not witness a significant increase in R squared when adding controls, so we do not choose to include them in fitting the index.

Table A.15: Political incentives and engagement in experimentation

	Engaged in experimentation			
	(1)	(2)	(3)	(4)
GDP per capita	0.021***	0.009***	0.045***	0.026***
	(0.001)	(0.001)	(0.003)	(0.002)
# of obs.	68,335	70237	68,335	70237
Mean of DV	0.023	0.023	0.023	0.023
Controls for political distortion	No	Yes	No	Yes
Policy FE	No	No	Yes	Yes

Note: The robust standard errors for (columns 1 & 2), and standard errors clustered at policy level (columns 3 & 4) are reported below the estimates. The purpose of the exercise is to account for the magnitude of positive selection due to misaligned incentives. The controls for political distortion include the career incentives of prefecture party leader, its interaction term with the hierarchical level of the city leader, and the indicator for whether a prefecture is enjoying political patronage (as described in Section 4.4). This analysis is carried out in a subsample of experiments targeting prefectural cities only since all political distortions we observed are at the prefectural level.

References

Bo, Shiyu. 2020. "Centralization and regional development: Evidence from a political hierarchy reform to create cities in china." *Journal of Urban Economics* 115:103182.

Cui, Jingbo, Junjie Zhang, and Yang Zheng. 2021. "The Impacts of Carbon Pricing on Firm Competitiveness: Evidence from the Regional Carbon Market Pilots in China." *Available at SSRN 3801316.*

Li, Pei, Yi Lu, and Jin Wang. 2016. "Does flattening government improve economic performance? Evidence from China." *Journal of Development Economics* 123:18–37.

Miratrix, Luke W, Jasjeet S Sekhon, and Bin Yu. 2013. "Adjusting treatment effect estimates by post-stratification in randomized experiments." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (2): 369–396.

Wang, Shaoda. 2016. "Fiscal competition and coordination: Evidence from China." *Department of Agricultural and Resource Economics, UC Berkeley, Working Paper.*

Yu, Jinkai, and Jing Yu. 2020. "Evolution of mariculture insurance policies in China: Review, challenges, and recommendations." *Reviews in Fisheries Science & Aquaculture,* 1–16.