Policy Experimentation in China: the Political Economy of Policy Learning

Shaoda Wang
David Y. Yang*

October 16, 2021

Abstract

Many governments have engaged in policy experimentation in various forms to resolve uncertainty and facilitate learning. However, little is understood about the characteristics of policy experimentation, and how the structure of experimentation may affect policy learning and policy outcomes. We aim to describe and understand China's policy experimentation since 1980, among the largest and most systematic in recent history. We collect comprehensive data on policy experimentation conducted in China over the past four decades. We find three main results. First, more than 80% of the experiments exhibit positive sample selection in terms of a locality's economic development, and much of this can be attributed to misaligned incentives across political hierarchies. Second, local politicians allocate more resources to ensure the experiments' success, and such effort is not replicable when policies roll out to the entire country. Third, the presence of sample selection and strategic effort is not fully accounted for by the central government, thus affecting policy learning and distorting national policies originating from the experimentation. Taken together, these results suggest that while China's bureaucratic and institutional conditions make policy experimentation at such scale possible, the complex political environments can also limit the scope and bias the direction of policy learning.

*. Wang: University of Chicago; shaoda@uchicago.edu. Yang: Harvard University, NBER and CIFAR; davidyang@fas.harvard.edu. We thank Daron Acemoglu, Isaiah Andrews, Abhijit Banerjee, Tim Besley, Mike Callen, Esther Duflo, Rema Hanna, Max Kasy, Ben Olken, Gerard Padro, Rohini Pande, Gautam Rao, Gerard Roland, Jesse Shapiro, Michael Song, Andrei Shleifer, Jaya Wen, Yang Xie, and Daniel Xu for stimulating comments and suggestions. Weicheng Cai, Qingyu Chen, Andrew Kao, Kaicheng Luo, Jiarui Qian, and Bobing Qiu provided outstanding research assistance.

1 Introduction

Determining which policies to implement and how to implement them is an essential government task (e.g., Hayek 1978; North et al. 1990). Policy learning is challenging, as policy effectiveness often hinges on the nature of the policy, its implementation, the degree of tailoring to local conditions, and the efforts and incentives of local politicians to make the policy work.

Many governments have explicitly or implicitly engaged in policy experimentation in various forms in order to resolve policy uncertainty and to facilitate policy learning (e.g., Roland 2000; Mukand and Rodrik 2005). Sophisticated policy experimentation has ranged from sequences of trials and errors to rigorous randomized control trials in subregions of a country. Few, however, can compare to the systematic policy experimentation in China in terms of its breadth, depth, and duration. Since the 1980s, the Chinese government has been systematically trying out different policies across regions and often over multiple waves before deciding whether to roll out the policies to the entire nation.

This project aims to describe and understand China's policy experimentation since the 1980s. Many scholars have argued that the pursuit of extensive, continuous, and institutionalized policy experimentation was a critical mechanism that led to China's economic rise over the past four decades (e.g., Rawski 1995; Cao, Qian, and Weingast 1999; Roland 2000; Qian 2002). Nonetheless, surprisingly little is understood about the characteristics of policy experimentation, or how the structure of experimentation may affect policy learning and policy outcomes.

We focus on two characteristics of policy experimentation that may determine whether it provides informative and accurate signals on general policy effectiveness (Al-Ubaydli, List, and Suskind 2019). First, to the extent that policy effects are often heterogeneous across localities, representative selection of experimentation sites is critical to ensure unbiased learning of the policy's average effects. Second, to the extent that the efforts of the key actors (such as local politicians) can play important roles in shaping policy outcomes, experiments that induce excessive efforts through local political incentives can result in exaggerated signals of policy effectiveness.

We ask three questions. First, has the sample selection in China's policy experiments been representative? Second, do policy experiments create additional incentives and induce extra effort that are not replicable outside of the experimentation? Third, how do the non-representative sample selection and non-representative experimental situation affect government's policy learning and shape national policy outcomes?

To answer these questions, we collect comprehensive data on policy experimentation

in China between 1980 and 2020. Based on 19,812 government documents, we construct a database of 633 policy experiments initiated by 98 central ministries and commissions. For each policy experiment, we link the central government document that outlines the overall experimentation guidelines with all corresponding local government documents to record its local implementation, and we trace its roll-out across the country. We measure a variety of characteristics of policy experiments based on the associated government documents and other linked datasets, including ex-ante uncertainty about policy effectiveness, career trajectories of central and local politicians involved in the experiment, the bureaucratic structure of the policy-initiating ministries, the degree of differentiation in policy implementation across local governments, and local socioeconomic conditions.

We begin by investigating the selection of experimentation sites. The ability to learn from a balanced, representative sample is a primary goal for the central government, as prescribed by the the National Development and Reform Commission, which oversees many key experiments. Nonetheless, comparing the pre-experimentation characteristics of the locations that are selected as test sites and those that are not, we observe that more than 80% of the experiments were conducted in sites that are positively selected in terms of local economic conditions. Such deviation from representativeness cannot be fully justified by optimal experimentation considerations. Rather, we document that nearly half of the observed positive selection can be accounted for by misaligned incentives across political hierarchies. Specifically, the level of promotion incentives faced by local politicians (which are greater for politicians who are sufficiently far away from retirement and for those who have ample room for upward mobility) shape their participation in the experiments, and political patronage affects how ministers choose experimentation sites.

Next, we examine whether policy experimentation induces politicians' strategic efforts during experiments, thus generating non-representative experimental situations. Using a triple-differences strategy, we find that during experimentation, local governments spend almost 5% more funds in the domains relevant to the policy on trial; this is particularly the case for politicians facing stronger promotion incentives. Such an increase in fiscal support is absent when the policy rolls out to the entire country. Moreover, we find that, among local politicians participating in a specific policy experiment, those facing greater career incentives act significantly differently in terms of policy implementation than those politicians who are not facing such strong career incentives. Such differentiation and potential recognition by the central government could earn local politicians substantial political credits.

Finally, we investigate whether the presence of positive selection in experimentation sites and local politicians' strategic efforts during experimentation affect the central gov-

ernment's policy learning and the national policy outcomes. We present evidence that the central government does not fully account for sample selection and strategic effort when evaluating policy experimentation. Experiments conducted in positively selected sites are substantially more likely to be promoted to national policies. When experimentation sites experienced exogenous positive shocks in fiscal resources (due to unexpected land revenue windfalls during the experimentation) or political incentives (due to local politician turnover occurring during the experimentation), the policies on trial are significantly more likely to be rolled out as national policies despite the fact that the innate effectiveness of these policies is orthogonal to those shocks. Furthermore, we find that evaluations of experimentation outcomes in the presence of positive sample selection and non-representative experimental situation can influence national policy outcomes. When the trial policies are rolled out to the entire country, localities benefit substantially more from the policies if they share similar socioeconomic conditions or comparable local politicians' career incentives with the corresponding experimentation sites. This could systematically bias the effectiveness of reforms in China, and generate distributional consequences across regions.

Taken together, these results highlight that China's remarkable policy experiments, as with any other undertaking in policy learning at this scale, take place in complex political and institutional contexts. On the one hand, certain institutional and bureaucratic conditions may serve as the engine to coordinate experimentation, to motivate politicians' participation, and to stimulate local policy innovations. Experimentation thus can help circumvent political and bureaucratic frictions that may prevent reform and policy adoption. On the other hand, as our results suggest, the very same institutional and bureaucratic contexts also imply the presence of factors that could result in deviation from representativeness in both sample selection and experimental situation. If these characteristics of the policy experiments are not sufficiently accounted for, policy learning can be biased and national policy outcomes may be affected.

This paper brings an important data point to the largely theoretical literature on policy learning and policy experimentation. For example, Aghion et al. (1991) and Callander (2011) provide theoretical frameworks on searching for good policies through experimentation; Dewatripont and Roland (1995) provide justification for the experimentation approach in policy reforms; Qian, Roland, and Xu (2006) study the relationship between government organizational structure and experimentation behavior; Hirsch (2016) analyzes experimentation in political contexts, where the objectives of learning and persuasion across decision-makers are intertwined; and Callander and Harstad (2015) investigate how decentralized jurisdictions strategically engage in policy experimentation, and

how a central government could help encourage policy convergence. Closest to the context we study, Montinola, Qian, and Weingast (1995), Cao, Qian, and Weingast (1999), Heilmann (2008a, 2008b), and Xie and Xie (2017) study the institutional setup and political logic of China's policy experimentation. We contribute to this body of work by linking the theoretical predictions on when policy experiments should take place and how they should be structured, with the first empirical analyses of the comprehensive set of policy experiments that have been conducted in China over the past four decades. We highlight that specific institutional contexts inevitably affect the structure of experiments and shape their outcomes.

Our work also joins the recent literature on policy learning and policy scale-up. Several studies highlight the structural factors that may limit how policy trials can inform broader outcomes after pilot programs are scaled up (e.g., Davis et al. 2017; Al-Ubaydli, List, and Suskind 2019). The patterns we document — positive experimentation sites selection in general, and, in particular, the diminishing policy effects as the policy is expanded beyond the site of better socioeconomic conditions and extra political incentives — echo the similar findings by Allcott (2015) on the sample selection bias in the Opower energy conservation programs, as well as findings by DellaVigna and Linos (2020) that trials conducted by the Nudge Units had smaller effects when scaled up due to changes in the intervention, institutional contexts, and implementation details. Our finding is also consistent with the prediction by Al-Ubaydli, List, and Suskind 2019 that competition among researchers (in our context, local politicians) could exacerbate the signal biases. Intriguingly, these patterns stand in contrast with the limited positive selection among the US states leading the policy innovations (DellaVigna and Kim 2021) and limited site selection bias in conditional cash transfer and microcredit experiments initiated by the Jameel Poverty Action Lab or Innovations for Poverty Action (Gechter and Meager 2021).[1]

Moreover, as we document that the Chinese government at times fails to disentangle factors not associated with inherent policy effectiveness when evaluating outcomes of policy experimentations, we join a number of recent studies in demonstrating that learning from policy trials may be further affected by decision-makers who are not sophisticated at processing information. They may not internalize information acquisition costs due to political hierarchy (Rogger and Somani 2018), take into account the context

1. Recent work also emphasizes the limits of local policy trials due to the general equilibrium consequences arising from policy scaling up (e.g., Bergquist et al. 2019), and factors related to external validity more generally (Vivalt 2020). Considerations of the external validity of experimental design have been central to much of the discussion, though it is typically focused on individual participants in the policy interventions and experiments, rather than on the localities (e.g., Snowberg and Yariv 2018).

of the study (Hjort et al. 2019), or consider the uncertainty of statistical inference (Vivalt and Coville 2019). Interestingly, Mehmood, Naseer, and Chen (2021) find that training on causal inference could increase policymakers' demand for and responsiveness to causal evidence on policy effectiveness.

The rest of the paper is organized as follows. Section 2 provides institutional background on China's policy experimentation. Section 3 describes the data sources, the process of constructing the database on policy experimentation, and a number of key characteristics on policy experimentation. Section 4 presents results regarding sample selection of experimentation sites. Section 5 presents results on strategic efforts by local politicians during the experiments. Section 6 presents evidence on the consequences of sample selection, strategic efforts and shocks on policy learning and national policy outcomes. Finally, Section 7 concludes.

2 Institutional background

China's policy experimentation represents a process "in which experimenting units try out a variety of methods and processes to find imaginative solutions to predefined tasks or to new challenges that emerge during experimental activity" (Heilmann 2008b).

The central government plays a key role in initiating and coordinating policy experimentation. While China's economic reforms are often accompanied by decentralization, high-powered political centralization remains a key characteristic of China's policy evolution (Xu 2011). It is thus important to note that China's policy experiments are not freewheeling trial and error or spontaneous policy diffusion. They are "experimentation under hierarchy," specifically, "purposeful and coordinated activity geared to producing novel policy options that are injected into official policy-making and then replicated on a larger scale, or even formally incorporated into national law" (Heilmann 2008b). Such a top-down approach to policy experimentation stands in contrast to the spontaneous experiments that often take place in federalist polities (Shipan and Volden 2006; Cai, Treisman, et al. 2009; Callander and Harstad 2015). While the policy experiments in China often begin with a small set of local governments, if the initiatives are deemed worth pursuing, they quickly move up the political hierarchy and enter a formal experimentation stage (if the central government chooses not to immediately make them national policies).

China's (and the Chinese Communist Party's) tradition of policy experimentation can be traced back to the Communist Revolution during the 1940s, most notably through the sequenced implementation of land reform in selected regions in order to consolidate the Communist regime. Interestingly, such policy experiments were driven primarily by the

lack of state capacity — policies as complicated as the land reform simply could not be implemented simultaneously and in a uniform manner across all regions under the Communist rule. The Communist Party took advantage of this policy implementation process, continuously adapting and tailoring policies as they were rolled out across localities. This became the earliest form of the "from points to surface" characteristic that defines China's policy experimentation.

Conducting policy experimentation before adopting the policies nationwide was institutionalized by Deng Xiaoping and Chen Yun in the 1980s and 1990s as a core principle guiding the reform and opening-up era policy transitions (Heilmann 2008a; Xie and Xie 2017). While the policy experiments during the Communist Revolution and early years of the People's Republic of China typically involved pre-conceived, centrally-imposed model emulation, the policy experiments during the Reform and Opening-up era are distinguished by their open-endedness in generating novel policy instruments and policy solutions. The "institutional entrepreneurship" released by policy experimentation has long been regarded as a key factor ensuring the stable deepening of China's economic reforms (Naughton 1996).

Primary form of experimentation: *experimentation points* The most pervasive form of policy experimentation in China is the selection of "experimentation points" (*Shidian*), as noted by Heilmann (2008a, 2008b). Before deciding whether a new policy should be implemented nationwide, the central government first tries out the policy regionally in a limited number of sites, possibly repeating the experiment in several waves, in order to evaluate the costs and benefits of the policy. Such a gradual approach allows effective policy innovations to precede "from point to surface," which could help avoid costly mistakes at the national level.

Heilmann 2008b describes China's policy experiments in general, and experimentation points in particular, as an inherently political process:

> [T]he effectiveness of experimentation is not based on all-out decentralization and spontaneous diffusion of policy innovations. China's experiment-based policy making requires the authority of a central leadership that encourages and protects broad-based local initiative and filters out generalizable lessons but at the same time contains the centrifugal forces that necessarily come up with this type of policy process.

The central government generally announces and introduces the policy experiments by publishing general guidelines. Such documents are issued by the ministries and commissions that lead the experiments, sometimes co-signed by coordinating ministries or

the State Council if inter-ministerial coordination is involved. The local government of each experimentation site typically responds to the central government documents by publishing a local experimentation action plan, laying out logistical and implementation details for the experiment.

The central government usually directly assigns certain regions as sites for experiments, but sometimes solicits local governments that would be willing to participate (Zhou 2013). Typically, the central government chooses experimentation sites at the province level, and then the provincial governments further delegate the experimentation to specific prefectural cities or counties within their jurisdictions.

A subset of the policy experiments is clustered in "experimental zones" (*Shiyanqu*). These are regions selected by the central government and given broad discretionary powers to try out various new policy bundles, essentially "creating a new system alongside, or in the interstices of, the existing one" (Naughton 1996).[2]

Once a policy experiment is determined to be successful, certain experimentation points are set as "demonstrational zones (*Shifanqu*)," and their experience in implementing the new policy will be actively promoted by the central government to the rest of the country (hence the term, "from point to surface"). Effective policies based on the experiments eventually are formalized by the central government and become national policies. In contrast, if a policy experiment fails to generate desirable outcomes — whether due to the policy's inherent ineffectiveness, local political economy constraints, high implementation cost, or unexpected public pressure against its implementation — the policy experimentation quietly stops expanding beyond the initial implementation stage. Few failed policy experiments are explicitly canceled.

In this paper, we focus primarily on policy experiments through experimentation points, including those clustered in experimentation zones. Most major reform initiatives in post-Mao China have been tried out by means of experimentation points before they were rolled out to the entire country (if at all); Appendix A.1 describes several other, less common forms of policy experimentation in China. Notable examples of policy experimentation through experimentation points in recent decades include reforms in local fiscal empowerment (2002 - 2015), carbon emission trading (2011 - 2021), separation of permits and licenses (2015 - 2018), and introduction of agriculture catastrophe insurance (2017 - 2021). We will describe these experiments in greater detail in Section 3.3.

2. The purpose of the experimental zones is to explore integrated bundles of economic development policies, rather than to evaluate the effectiveness of a specific policy, which is conceptually closer to Sachs (2006). The most notable examples for experimental zones are the Shenzhen Special Economic Zone and Shanghai Pudong Special Economic Zone, which have served as policy laboratories for various reforms during the Reform-and-Opening era.

3 Data and characteristics of policy experimentations

We compile, to the best of our knowledge, the most comprehensive dataset on policy experimentation in China over the past four decades. Our primary data source relies on official government documents, which we describe in Section 3.1. We also complement the government documents with a number of auxiliary datasets such as local socioeconomic conditions and the background of involved politicians; we describe these data sources in Appendix B. We present, in Section 3.2, a number of characteristics of the policy experiments that we construct based on the government documents and auxiliary datasets. We illustrate four policy experiments as stylized examples in Section 3.3.

3.1 Government documents on policy experimentation

Our main data is based on the comprehensive collection of policy documents issued by the Chinese central and local governments since 1949 compiled by *PKULaw.com*, an online platform hosted by Peking University Law School.

Specifically, we collect (nearly) the universe of government documents between 1980 and 2020 containing the key words "experimentation points" (*Shidian*) and "experimental zones" (*Shiyanqu*). We obtain 19,812 documents in total, among which 4,399 were issued by the central government and 15,413 by local governments. Central government documents mark the official initiation of particular policy experiments, their key milestones (e.g., when a major expansion of experimentation is planned), and decisions to roll out the policies to the entire country if the experiment is successful. Local government documents are issued by each locality participating in the experiments, specifying details on local implementations and administrative arrangements.

We identify 633 distinct policy experiments based on policy themes. Our categorization of policy experiments is conservative: consecutive experiments are grouped into the same policy experiment as long as they concern similar policy aims, even if the specific contents of the policies evolve and even if the names of the policies change. Moreover, policy experiments that are closely related and simultaneous in implementation are combined into one experiment, even if the central government issued separate documents for each component.[3] We distinguish different phases of the experiments by distinct waves of the experimentation roll-out, often marked by specific central government documents.

3. For example, experiments on corn seed insurance, rice production insurance, professional farmer training, and agricultural tech promotion consulting service are combined into an overarching experiment on improving agricultural technology and management.

Among the 633 policy experiments, 594 involve policies explicitly intended for potential national roll-out, and 39 are policies with specific regional targets.[4] For the baseline analysis in Section 4 where we examine experimentation sites selection, we exclude policies with explicit regional targets, though the results are robust if we include all policy experiments in the analyses and adjust the corresponding comparison of experimentation sites selection based on scope. Around 104 of the policy experiments are ongoing, and we will exclude them from the analyses in Section 6 where we examine whether the policies on trial roll out to the entire country.

Coverage of policy experimentation Initiation of experimentation from inside the government is by far the most frequent starting point (Heilmann 2008b). Government-initiated experiments have corresponding government documents, ensuring our comprehensive coverage on such experiments. In particular, our data includes extensive coverage on potentially failed experiments, as well as government documents that are expired, void, or explicitly revoked.

We conduct various cross-checks to ensure the comprehensiveness of the government documents that we collect. Specifically, for the ministries that publish documents on their own websites, we independently collect documents from the ministerial websites. We find that the government documents collected from the *PKULaw.com* has extensive and comprehensive coverage (see Appendix Table A.1). When we manually examine the limited documents that are published on the ministries' websites but not included in the *PKULaw.com* database, we find that they are secondary documents and do not contain information on policy experiments.

Because we are relying on government documents to describe policy experiments, the experiments must have reached a stage of formal endorsement and coordination by the central government in order to be included in our sample.[5] Thus, we do not observe very early stage experiments initiated by the local governments that never reach the level of the central government — e.g., early bottom-up policy entrepreneurship led by specific local governments that fail to receive the central government's approval for continuing and expanding the policy. This implies that the set of centrally coordinated policy experiments that we study is already positively selected in terms of the central government's prior evaluation of the policy's effectiveness. However, such sample selection does *not*

4. Examples of regional target policies are: anti-poverty policies aimed at rural regions, Chinese language education policy aimed at regions with a high share of ethnic minority population, industrial restructuring policy for the Northeast region, and free trade zone trials targeted at a few major ports such as Shanghai.

5. Promising policy innovations initiated by the local government escalate to the central government fairly rapidly, typically within a year or two after the first instance of the local policy trials.

mean that policy uncertainty is irrelevant in this context: on average, 46% of the policy experiments fail to become national policies, even though the central government envisioned all of them as having relatively high promise at the onset.

3.2 Characteristics of policy experiments

We extract several key pieces of information from the corresponding government documents in order to characterize each policy experiment.

Time of initiation We first extract information on the year when policy experiments are initiated. Figure 2 plots the number of experiments initiated in each year across the past four decades, where we record the first year when a specific policy experiment started as the year associated with the multi-year roll-out of the experimentation. We observe a hump-shaped pattern: the number of policy experiments initiated by the central government remained relatively low throughout the decades of the 1980s and 1990s, averaging less than 10 new experiments per year across all ministries and commissions. The number of experiments began to sharply increase toward the end of the 1990s, reaching a peak of 76 new experiments initiated in 2013 alone. Since 2013, the number of new experiments started to decline, and nearly halved by the end of our sample period in 2020.

While many factors could contribute to these patterns, at least part of the decline in the number of experiments in the recent decade can be attributed to the vertical management transition of many state ministries. As these ministries shift the control over their personnel, funding, and decision rights from the local governments to the upper-level ministerial units, they move away from flat, multi-divisional structures (M-form) that may provide flexibility and ease in coordinating policy experiments, to more centralized, unitary structures (U-form) that benefit from economies of scale. Consistent with the theoretical predictions (e.g., Chandler 1962; Williamson 1975; Qian, Roland, and Xu 2006), we find that, following the transition to U-form organization, the vertically managed ministries significantly decreased the number of policy experiments that they administer. Appendix C presents results using an event study design.

Experimentation sites and the roll-out schedule We extract the experimentation sites and the roll-out schedule of each policy experiment. Many policy experiments have more than one wave of roll-out, and we identify 1,374 distinct rounds of roll-out across the 633 experiments. We link each government document to a specific round of experimentation, which allows us to observe the time localities join a particular policy experiment and to compare the selection pattern of experimentation sites across rounds.

Figure 1, Panel A plots the distribution of experimentation sites across China, aggregated at the province level (see Appendix Figure A.1 for county level distribution). Table 1, Panel A presents the total number of policy experiments initiated during 1980 and 2020 and the average number of rounds and experimentation sites involved in each experimentation. On average, each policy experiment initiated by the central government contains more than two rounds in its roll-out and lasts for 2.25 years, until either the roll-out stops or the experiment becomes a national policy.

Policy domains and involved ministries We identify all the central government ministries and commissions involved in a policy experiment, and measure each ministry or commission's role in that experiment (e.g., initiator or collaborator). In cases where a particular policy experiment is introduced by multiple ministries and commissions, we also identify the primary ministry or commission that takes the leading role. A total of 98 ministries and commissions are involved, ranging from the State Council to the Ministry of Agriculture and the Ministry of Finance. Table 1, Panel B presents the number of policy experiments initiated by different ministries and commissions, grouped by policy domains and broad functions for which they are responsible. Appendix Figure A.2 plots the count of policy experiments by policy domain over time.

National roll-out We observe whether policy experiments are rolled out to the entire country and become national policies. This is marked by specific central government documents concluding the experimentation cycle. Overall, 53.9% of the policy experiments eventually became national policies, while 46.1% failed (see Figure 2, share of successful and failed experiments indicated by darker and lighter gray shades, respectively). The share of policy experimentation leading to national policy roll-out remains remarkably stable over time (see Appendix Figure A.3). The patterns concerning successful and failed policies are not sensitive to the particular definition; for example, we alternatively define a policy experiment as successful if the roll-out covers at least two-thirds of the whole country's counties, and we find similar patterns throughout (see Appendix Figure A.4).

Importance, complexity, and uncertainty We measure the importance, complexity, and ex-ante uncertainty of each policy experiment. We capture the degree of importance by whether an experiment is explicitly mentioned in the central government's *Five Year Plans*, which represent the most important policy blueprints issued by the Chinese government and cover policy agendas considered as of highest priority in the upcoming five-year period. 19.2% of the policy experiments reflect policy themes mentioned in the *Five Year*

Plans (see Appendix Figure A.5). We capture the degree of complexity by the number of ministries and commissions involved in the experiment, as well as the length of the initial documents describing the experiment. 24.3% of the policy experiments involve more than two ministries and commissions; we label these as complex experiments (see Appendix Figure A.6). We capture the degree of ex-ante uncertainty of policy experimentation based on whether the central government has laid out a detailed national roll-out timeline before the experiment starts. 30.7% of the experiments feature such timelines (which we label as experiments on policies with high certainty), and 62.0% of them eventually become national policies. In contrast, among the 69.3% of experiments that do not feature such a timeline (which we label as experiments on policies with high uncertainty), only 36.2% were eventually rolled out to the entire country (see Appendix Figure A.7).

Assigned vs. voluntary participation We categorize policy experiments as either assigned or voluntary, depending on whether the experimentation sites are designated and assigned by the central government directly, or the experiment invites voluntary participation of the local governments. About 40.0% of the experiments allow (at least partially) for voluntary participation of the local governments (see Appendix Figure A.8).

Auxiliary characteristics Finally, we measure a number of auxiliary characteristics of policy experiments, which we incorporate into various parts of the analyses. For example, we identify whether the central government would provide additional fiscal support for the experimentation sites, and whether the policies on trial would in principle benefit from extra fiscal support. These characteristics will help us evaluate the plausibly strategic fiscal resources allocated by the local governments in order to improve the local outcomes of the experiments. We also measure how policy innovation and differentiation evolve across time and space, by constructing matrices of pairwise textual similarities for all the local policy documents that belong to the same policy experiment.[6] Such measurement allows us to investigate the conditions under which local governments exert greater efforts to differentiate their local policy implementation.

3.3 Four examples of policy experimentation

We map four distinct policy experiments to illustrate the ranges of policy experimentation that took place in recent decades (see Appendix A.2 for additional details).

6. Text similarity is calculated using Latent Similarity Analysis, a canonical choice in natural language processing. After removing stop words, we conduct the TF-IDF encoding for each word vector, and then use the first three principal components to compute cosine similarity.

Figure 1, Panel B.1 depicts the experimentation on carbon emission trading policy initiated in 2011, which involves five prefectures (Beijing, Tianjin, Shanghai, Shenzhen, and Chongqing) and two provinces (Guangdong and Hubei), all of which are among the most developed localities in the country. The policy rolled out to the entire country in 2021, after just one wave of experimentation. Panel B.2 depicts the experimentation on a policy that aims at reducing administrative burdens to firm entry by separating permits from licenses for new firms: since 2015, the experiment has taken place among 24 prefectures over three waves, very much concentrated in the developed, coastal regions and provincial capitals. This policy rolled out to the entire country in 2018.

Panels B.3 and B.4 describe two experiments that did not lead to national policies. The experimentation on the introduction of agriculture catastrophe insurance started in 2017, and a total of 14 provinces participated as experimentation sites over two waves (see Panel B.3). These experimentation sites are inland provinces in Eastern China, as well as those in the Northeast. The experimentation ended after two waves and this policy did not roll out to the entire country. Finally, as depicted in Panel B.4, the experimentation on county fiscal empowerment reform took place over more than a decade, involving 1,246 counties as experimentation sites across more than 10 waves. The experimentation started with developed regions in the earlier waves and moved towards inland, less developed regions. The experimentation ended in 2015 and the fiscal empowerment reform did not roll out to the country.

4 Is the selection of experimentation sites representative?

Focusing on the policy experiments that are meant to test potential national policies, we first examine which localities are selected as experimentation sites. As a benchmark, we examine whether the selection of experimentation sites is indeed representative.

From the central government's perspective, a key criterion for experimentation site selection is its representativeness, which determines the quality of knowledge one could extract from a policy experiment (Zhou 2013). The National Development and Reform Commission, the leading governance body that guides and coordinates national policies, lays out the overall principles of choosing experimentation sites as:

> The balanced distribution of experimentation sites is the most important criteria in choosing these sites. [...] Policy experiments are not meant to solve development problems of a particular place or a particular sector. Rather, they need to gather knowledge and experiences for the policy reform and institu-

tional innovation at the national level. [...] Hence, the experimentation sites should be fairly representative.

4.1 Procedure to test for representativeness

For each policy experiment, we compare pre-experimentation characteristics between locations where the experiment is implemented and those that do not participate in the experiment. As the baseline, we examine the local GDP per capita in the year prior to experimentation roll-out.

We conduct t-tests against the null hypothesis that the pre-experimentation levels of local GDP per capita are indistinguishable among the experimentation sites and non-experimentation sites. This amounts to 633 independent t-tests, one for each policy experimentation.[7] Note that conducting representativeness tests separately for each policy experiment is conservative: if one were to identify deviations from representativeness with these separate tests, then a pooled test with multiple experiments would yield more power in detecting unrepresentativeness and rejecting the null hypothesis. We discuss below the robustness of using a variety of other local characteristics as well as alternative testing methods.

We use the corresponding t-statistics as summary statistics to quantify the deviation from representativeness for each policy experiment. Specifically, the *studentized-t* statistic for policy experiment i is:

$$t_i = \frac{\hat{Y}_i(1) - \hat{Y}_i(0)}{\sqrt{\frac{\hat{S}_i^2(1)}{n_{i,1}} + \frac{\hat{S}_i^2(0)}{n_{i,0}}}}, \tag{1}$$

following the t-distribution with degrees of freedom ν_i, where

$$\nu_i = \frac{\left(\frac{s_{i,1}^2}{n_{i,1}} + \frac{s_{i,2}^2}{n_{i,2}}\right)^2}{\frac{\left(s_{i,1}^2/n_{i,1}\right)^2}{n_{i,1}-1} + \frac{\left(s_{i,2}^2/n_{i,2}\right)^2}{n_{i,2}-1}}. \tag{2}$$

The specific context of China's policy experimentation poses two important complications in conducting the representativeness tests. First, policy experiments can be implemented at the provincial level, prefectural level, or county level.[8] The county and pre-

7. For each policy experiment's representativeness test, we adjust the respective degree of freedom in the underlying distribution based on the exact share of localities that participate in the experiment.

8. Centrally-administered municipalities are considered as either provinces or prefectures, depending on the level of policy experimentation implementation. As we discuss below, our baseline patterns remain robust if we exclude these municipalities from the analyses.

fectural level experimentations often represent cases where experimentation provinces are selected by the central government, and the corresponding provincial governments then choose the counties or prefectures within their jurisdictions to implement the experiment. Thus, the experimentation sites selection has distinct administrative samples. We conduct the representativeness tests at the appropriate level for each policy experiment. For county and prefectural level experiments, the tests are conducted at the corresponding county or prefectural level, stratified based on the experiment-participating provinces — in other words, counties or prefectures participating in the experiment are compared only with other non-experimenting counties or prefectures within the same province.

Second, approximately one-fourth of the experiments involve only one experimentation site. We cannot conduct standard statistical tests for these one-site experiments. Instead, we pool each one-site experiment with four other randomly selected one-site experiments, and conduct the representativeness test on the pooled sample, where the non-experimentation sites are defined as those that do not participate in any of the five experiments. This will yield a corresponding t-statistic for each of the one-site experiments. We conduct a range of alternative test specifications concerning these one-site experiments, such as pooling experiments that take place in consecutive periods, and drawing bootstrap samples with replacement.

4.2 Most experimentation sites are positively selected

We plot, in Figure 3, Panel A, the distribution of the t-statistics on comparing pre-experimentation local GDP per capita between the experimentation and non-experimentation sites. We mark the thresholds of t-statistics where one would reject the null hypothesis of representative experimentation site selection at the 90% confidence interval.[9] Table 1 reports the corresponding test statistics (adjusting for the degree of freedom for each test) and the share of policy experiments for which we can reject the null hypotheses at the 10% level.

We find that the experimentation sites for 80.7% of the experiments are richer on average (in terms of pre-experimentation local GDP per capita) than localities that do not participate in the corresponding experiment. Even with statistical tests that are fairly conservative, we are able to reject the null hypothesis of representative selection among 50.5% of the experiments at the 10% level.[10]

9. As discussed above, each of the 633 *t*-tests has its specific degree of freedom. We depict visually the average width of the 90% confidence interval (2.36).

10. We observe modest decrease in the positive selection of experimentation sites over the years, suggesting potential learning and adjusting by the central government. Appendix Figure A.9 plots the overall share of positively selected experiments over the four decades since 1980, and Appendix Table A.2 presents regression results on the time trend in experimentation's positive selection, for all experiments and separately

The pattern of positive selection of experimentation sites is remarkably robust. First, we observe similar patterns of positive selection when conducting tests using alternative regional pre-experimentation characteristics such as total local GDP, local population size, and local fiscal revenue (Figure 3, Panel B presents the summary statistics; see also Appendix Figure A.10, Panels A to C, respectively). Second, even stronger positive selection is observed when we zoom in to particular policy domains. For example, agricultural policy experiments take place in localities with substantially higher pre-experimentation agricultural output; experiments with government finance and tax policies take place in localities with substantially higher local fiscal revenue; and experiments with population and health policies take place in localities with substantially larger population (see Appendix Figure A.11).[11] Pooling all policies together and focusing on pre-experimentation fiscal expenditure in the policy-specific expenditure categories, we again find strong patterns of positive selection (see Appendix Figure A.10, Panel D). Third, the patterns are similar when we incorporate one-site experiments: the representative tests pooling consecutive experiments display similar patterns (see Appendix Figure A.13, Panel A). It is robust to a variety of ways in which we pool and specify random draws of the pooled tests among the one-site experiments (see Appendix Figure A.13, Panel B), and to alternative permutation tests among the multi-site experiments (see Appendix Figure A.13, Panel C). Fourth, the pattern of positive selection is robust if we examine just the subsample of early-round experimentation sites, effectively holding fixed the number of experimentation sites across policy experiments (see Appendix Figure A.14). Finally, the pattern of positive selection is robust if we exclude the selection of centrally-administered municipalities such as Beijing and Shanghai, where local economic development and central government's priority in policy implementation coincide (see Appendix Figure A.15).

4.3 Unlikely explanations of the observed positive selection

What may explain the positive selection of experimentation sites? We next document a number of stylized patterns that could help rule out certain explanations.

Ex-ante policy uncertainty One may speculate that, depending on the ex-ante uncertainty that the central government holds toward each policy on trial, the specific objectives of the experimentation could differ and thus justify the deviation from representa-

by ministry.

11. Taking into account of the fact that different policy domains have different unit of analyses — for example, estimating policy effects on firms may require positive selection on counties as experimentation sites — does not explain vast majority of the positive selected policy experiments that we observe (see Appendix Figure A.12).

tive sample selection. Experiments on policies that the central government is more certain about rolling out to the entire country (captured by whether the central government specifies a timeline for such national roll-out *before* the experiment starts) might not have learning about policy effectiveness as the primary goal. However, when we separately evaluate the degree of representativeness in site selection among experiments that are ex-ante certain and those that are ex-ante uncertain (see Table 1, Panel D), we find that the site selection bias among ex-ante uncertain policies is in fact substantially higher (average t-statistics = 2.95) than that among ex-ante certain policies (average t-statistics = 2.12).

Complex experiments Positive selection of experimentation sites could be justified if richer localities — often represented by better local governance and administrative capacity — may be better at carrying out the demanding trial policies and thus provide more precise signals on the policy effectiveness. Such justification for positive selection could be even stronger for complex experiments, for example, those that involve coordination and collaboration across multiple ministries and local government bodies. Nonetheless, as shown in Table 1, Panel E, we observe that the site selection among experiments that are less complex, involving a single ministry or commission, deviates (slightly) further from representative than those that are more complex, multi-ministerial experiments (average t-statistics = 2.84 vs. 2.65, respectively).

Eventual scope of policy roll-out Positive selection of experimentation sites could also be justified if the intended geographic scope of the eventual policy is limited to richer localities. While the vast majority of the policy experiments initiated by the central government concerns national policies, there exist different degrees of flexibility in regional targeting across policy domains. Table 1, Panel B presents the results of the representative tests for experimentation across policy domains. We observe that experiments on policy domains such as market supervision that are more likely to be nationally uniform are *more* positively selected (average t-statistics = 3.22) than domains such as agriculture that are more flexible in terms of sub-national targeting (average t-statistics = 1.98).

Optimal experimentation Unrepresentative roll-out of experimentation may be justified if the central government has other objectives in addition to learning about the true underlying treatment effects and persuading other agents who might hold different priors. To evaluate the importance of these alternative objectives, we conduct quantitative exercises to incorporate alternate objectives studied by the recent literature. Specifically, we examine the incentives of subjective expected utility, in addition to learning and per-

suasion, following Banerjee et al. 2020. We simulate the optimal experimentation design, parameterizing the model based on data from the 25th, 50th, and 75th percentile Chinese policy experiments in terms of their degree of positive selection. Appendix D.1 provides details on the simulation procedure.

We find that when the central government places heavier weight on its subjective expected utility, deterministic experimentation becomes more justified than randomization. However, even if one places 100% of the weight on the decision maker's subjective expected utility, only less than 5% of the optimal designs for these experiments induce positive selection with *t-stats* > 1, with the optimal *t-stat* never exceeding 2.6 — substantially lower than the positive selection that actually occurs.

We additionally test two extensions on the model presented: *(i)* allow for the experimental information (or equivalently, policy execution) quality to vary with local county's GDP; and *(ii)* allow counties to consent or opt-in to treatment so that only counties with positive treatment effects are treated. Although both extensions mildly increase selection, the *t-stats* from these simulations still remain much lower than those observed in reality.

4.4 Political sources of deviation from representative sample selection

Could positive selection occur even if the central government genuinely intends to conduct representative experimentation, as suggested by the National Development and Reform Commission? Does the central government have alternative goals or constraints that prevents it from executing representative sample selection? In this section, we investigate the political factors that lead to the sample selection.

Local politicians' career incentives We first examine how the prefectural leaders' incentives for career advancement affect their participation in policy experimentation.

A number of patterns suggest that local politicians' incentives to positively represent the results of policy experiments indeed play a role in generating positive site selection. First, on average, participation in successful policy experiments is associated with a 22.3% increase in promotion probability for the corresponding local politicians (see Appendix Table A.3). When local politicians are facing stronger career incentives in a certain year, they may have stronger motives to improve their portfolio of political achievements, including through participation in important and successful policy experiments (Wu 1995; Huang 2000). Second, we find that the deviation from representativeness is not nearly as severe at the province level, as compared to the choices of specific prefectures and counties to be the experimentation sites (see Table 1, Panel C). Third, experiments are closer to being representative if the site selection is assigned by the central government directly

rather than involving voluntary participation by the local government (see Table 1, Panel F).

To test this hypothesis more formally, we follow Wang, Zhang, and Zhou (2020) and estimate each prefectural city leader's *ex-ante* likelihood of promotion in each year, as a flexible function of their age (relative to retirement), tenure and official rank in the bureaucratic system (capturing the potential for upward mobility); Appendix B.1 provides details of the construction of this measure.

Then, we estimate the following econometric model by exploiting within-prefecture changes in leaders' political incentives:

$$y_{pt} = \alpha \cdot Incentive_{pt} + X'_{pt} \cdot \beta + \delta_p + \theta_t + \varepsilon_{pt}, \tag{3}$$

where y_{pt} is the number of policy experiments in prefectural city p in year t; $Incentive_{pt}$ is the estimated promotion incentive index for the political leader of region p in year t; and X'_{pt} is a vector of time-variant regional control variables. Importantly, we control for full sets of region fixed effects and year fixed effects (δ_p and θ_t, respectively), thus identifying the political incentive effects from within-prefecture, across-year discontinuous changes in career incentives, due either to politicians' aging and changes in their opportunities for promotion or to local leaders' routine turnover.

As shown in Table 2, Panel B, when the prefectural leaders have stronger promotion incentives, the corresponding localities engage in significantly more policy experiments. Reassuringly, we do not observe similar effects with the promotion incentives among the preceding politicians who should not have direct influence on subsequent engagement in policy experiments (see Appendix Table A.4). Moreover, such effects of promotion incentives are almost entirely driven by policy experiments initiated by M-form ministries (see in Appendix Table A.5). Since the U-form ministries are directly administered by the central government, the local politicians would have neither capacity nor incentives to influence experiments initiated by U-form ministries (as compared to those initiated by M-form ministries). This is because U-form initiatives are not under the jurisdiction of local governments, and, as a result, local politicians receive less credit for successful experimentation. This pattern also suggests that our findings are unlikely driven by omitted confounding factors: an omitted factor could confound our results only if it were correlated specifically with policy experiments initiated by M-form ministries.

Political patronage Misaligned incentives could also be present within the central government — between the policy experimentation coordination bodies such as the National Development and Reform Commission and the specific ministries in charge of the experi-

mentation. Given the potential political rewards associated with successful policy experimentation, political patronage — prevalent in China's political system (Fisman and Wang 2015; Fisman et al. 2020) — could also shape the selection of experimentation sites, due to reasons such as favor exchange, higher trust among political patriots, and ministers' better control over local implementation.

To investigate this hypothesis, we exploit the inter-temporal changes in a region's connection to each ministry caused by the turnover of ministers at the central government level. Specifically, we define a province as connected to a ministry if the current minister used to work full-time in that province before becoming the minister. To the extent that the local governments cannot influence the appointment of central ministers, the turnover of ministers can be regarded as exogenous shocks to the province-ministry connections.

We estimate the following econometric model using ministry-province-year level data:

$$y_{mpt} = \alpha \cdot Connection_{mpt} + X'_{pt} \cdot \beta + \delta_{mp} + \theta_t + \varepsilon_{mpt}, \tag{4}$$

where y_{mpt} is the number of experiments assigned to province p by ministry m in year t; $Connection_{mpt}$ is a dummy variable indicating whether the minister of ministry m in year t used to work full-time in province p; X'_{pt} is a vector of provincial time-variant controls; and θ_t is year fixed effects. Importantly, we include δ_{mp}, province-by-ministry fixed effects, which isolate the changes in a locality's connection to a particular ministry driven by minister turnovers.

As shown in Table 2, Panel A, when a region becomes connected to a minister, the number of experiments assigned to that region increases immediately by 28.8%.[12] The effects are almost entirely driven by cases where the central ministry directly assigns the experimentation sites, while there is no comparable effect when the experimentation sites are selected via voluntary participation (see Appendix Table A.6). This suggests that the political patronage in experimentation site selection works through top-down favoritism.

Objectives beyond representative sample selection The patterns on career incentives and political patronage suggest that, even in the case that the central government's only objective is to obtain a representative sample of experimentation sites, deviation from representativeness in site selection could still occur in the implementation of policy experiments: principle-agent problems generate misaligned incentives between the central and local governments, and between the central government and its ministers.

In addition to achieving representativeness, the central government might have other

12. In Appendix Figure A.16, we plot the event study estimates around ministers' turnover. The absence of a pre-trend suggests that being connected to a ministry due to turnover of a central minister is indeed likely to be orthogonal to the counterfactual trajectories of local governments' experimentation behaviors.

criteria when choosing experimentation sites, and these additional criteria may at times counteract its desire for representativeness. A particular example of such criteria concerns the central government's often overt and overarching objective of maintaining political stability during socioeconomic reforms. Specifically, we examine whether social and political unrest in a particular prefecture affects its chance of being selected as an experimentation site. We find a robust pattern, exploiting within-region, across-time variations in occurrences of unrest, showing that prefectures that have experienced social and political unrest in the preceding period are significantly and substantially less likely to become experimentation sites (see Table 2, Panel C). This suggests that unstable local environment could be a veto condition that precludes participation in policy experimentation.The negative relationship between unrest and selection is much stronger in the case of top-down assignment to experimentation than in the case of voluntary participation (see Appendix Table A.7). This suggests that concerns about avoiding politically unstable localities are primarily held by the central government, rather than by potential local participants.

Accounting for observed positive selection Overall, the factors associated with misalignment across the political hierarchy could account for nearly 50% of the positive selection in experimentation sites that we observe. We provide several quantitative assessments of these factors in contributing to site selection in Appendix E.

5 Do experiments induce strategic efforts?

An important component of policy effectiveness is an incentive scheme that encourages sufficient effort from the local governments when they implement the policy. A policy experiment — perhaps due to its high visibility, high political reward, and explicit monitoring by the central government — may induce additional efforts by the local governments who are especially incentivized to make the policy at trial appear successful at its experimentation stage. While participation in successful experiments is associated with a substantial increase in local politicians' promotion, this is not the case for participation in failed experiments (see Appendix Table A.3).

In this section, we examine two particular aspects of deviation from representative experimental situation: local governments' allocation of fiscal resources (Section 5.1) and their efforts to differentiate during the implementation of experiments (Section 5.2).

5.1 Allocation of fiscal resources during experimentation

Local fiscal expenditure is an important input in policy outcomes, and is strongly associated with overall local economic performance (see Appendix Table A.8). Do the local governments participating in policy experimentation significantly increase fiscal expenditure that may improve the outcome of the experiment?

To answer this question, we first match each policy experiment to one of the six broad fiscal expenditure domains that are consistently reported in the county fiscal expenditure data throughout our sample period: general administrative cost, infrastructure, economic production, agriculture/forestry/fishing, science/culture/education/health, and other. We then use a triple-differences strategy to examine whether the start of policy experimentation is associated with increases in fiscal expenditure in the corresponding domain. Specifically, we estimate the following econometric model using county-domain-year level data:

$$y_{ikt} = \alpha \cdot Exp_{ikt} + \lambda_{it} + \delta_{kt} + \theta_{ik} + \varepsilon_{ikt},$$

where y_{ikt} is the ratio of fiscal domain k specific to the experiment in the total fiscal expenditure in county i during year t; Exp_{ikt} is the number of experiments in fiscal domain k that county i engaged in during year t; λ_{it}, δ_{kt}, and δ_{kt} stand for county-by-year, domain-by-year, and county-by-domain fixed effects, respectively.

The results are presented in Table 3, Panel A. We observe a substantial increase in domain-specific fiscal expenditure (columns 1-3): an additional experiment increases the local expenditure in the corresponding category by about 2% in terms of share of total fiscal expenditure, and by more than 5% in terms of the level of expenditure. The increase in domain-specific fiscal expenditure during experimentation is greater if the local politicians face stronger career incentives at the time of the experiment (columns 4-6), consistent with the pattern that politically incentivized local leaders are particularly keen on making sure the policy experiments succeed in their regions of jurisdiction.

Importantly, the over-expenditure during the experimentation may not be sustained when a policy becomes national. Indeed, we do *not* find fiscal expenditure increases in specific domains among non-experimentation sites when the policy rolls out to the entire country, and this is the case regardless of career incentives of the local politicians at non-experimentation sites (see Table 3, Panel B).[13]

13. This finding echoes similar results that document short-term "window dressing" incentives among local politicians when their actions are more visible to the central government (e.g., Fang, Liu, and Zhou 2020).

5.2 Political incentives and differentiation during experimentation

Next, we examine whether local politicians with stronger career incentives differentiate their experimentation details more during policy implementation. Differentiation can signal effort and earn political credit as a "model experimentation site."

In order to capture local politicians' differentiation, we measure the extent to which local politicians issue policy experimentation documents that are distinct from the ones issued by other politicians participating in the same experiment. Specifically, we construct pairwise text similarity among documents issued by local governments on the corresponding policy experiment, calculated using Latent Similarity Analysis (LSA). This exercise follows Bertrand et al. (2020) and Acemoglu, Yang, and Zhou (2021) in spirit, and we describe details of the procedure in Appendix F.

After constructing pairwise text similarity across documents issued by the local governments for a specific experiment p, we measure each local government i's similarity with its peers that have participated in the same experiment in a previous wave, using the maximum similarity score among these pairs (y_{ip}). We estimate the following econometric model:

$$y_{ip} = \alpha \cdot Incentive_{ip} + \beta X'_{ip} + \lambda_i + \delta_p + \gamma_t + \varepsilon_{ip},$$

where $Incentive_{ip}$ is the politicians' career incentives as in Section 4.4; X'_{ip} is a set of controls for the politicians (educational attainment and career experience in the central government); λ_i is a full set of location fixed effects; δ_p is a full set of policy experiment fixed effects; and γ_t is a full set of year fixed effects. Similarly to the exercise in Section 4.4, we exploit variations in politicians' career incentives due to the timing of the experiments and their age relative to retirement.

The results are presented in Table 4. We observe that, when local politicians have strong career incentives, they tend to differentiate more relative to their colleagues in terms of implementation details. While we cannot conclude whether such differentiation is sub-optimal (e.g., if policy solutions that are proven effective had already been tried out by their peers in previous waves of experimentation), the increase in policy implementation differentiation reflects an increase in local politicians' efforts in implementing the policy on trial.

6 What are the consequences on policy learning and policy outcomes?

Having demonstrated that the sample selection bias and unrepresentative experimental situation are relevant in China's policy experimentation, we next examine the consequences of their presence. To the extent that the central government may not fully take into account the sample selection of experimentation sites, and the endogenous efforts local politicians exert that may not be sustainable after the experimentation stage, then policy learning could become biased — we investigate in the policy learning consequences in Section 6.1. To the extent that the central government's policy learning may be biased, this may shape the national policy outcomes in China — we investigate in the consequences on policy outcomes in Section 6.2.

6.1 Central government's policy learning

Policy learning and locality-specific shocks When evaluating experimentation outcomes, does the central government exclude locality-specific shocks that are orthogonal to the underlying policy effectiveness? In particular, does a local fiscal windfall during experimentation, which may improve local outcomes but unrelated to the innate effectiveness of the policy at trial, increase the likelihood that the central government evaluate the policy at trial to be successful?

We focus on land revenue (i.e., land conveyance fees) received by the county governments for converting agricultural land for non-agricultural use. This has been one of the most important sources of local fiscal revenue windfall in the 2000s (e.g., Han and Kung 2015). Following the empirical strategy of Chen and Kung (2016), we isolate the exogenous component of such land revenue windfall as a result of the interaction of two factors: *(i)* the amount of land in a county suitable for commercial and real estate development as determined by terrain; and *(ii)* exogenous time-varying demand shock driven by interest rates. We evaluate whether the land revenue increase due to these factors unrelated to policy experimentation and policy effectiveness *per se* may affect the chance that a policy experiment becomes successful and gets rolled out to the entire country.

Using a sample of all experimentation sites for each policy experiment, we estimate the following two-stage-least-squared specification:

$$\begin{aligned} Land\_revenue_{ipt} &= \alpha \cdot Suitability_i \times Interest_t + X'_{it}\beta + \delta_i + \gamma_t + \delta_m + \epsilon_{ipt} \\ y_p &= \mu \cdot \widehat{Land\_revenue}_{ipt} + X'_{it}\Gamma + \psi_i + \nu_t + \delta_m + \varepsilon_{ipmt}, \end{aligned}$$

where $Land\_revenue_{ipt}$ is the log level of land conversion revenue obtained by county i, served as an experimentation site for policy p, in year t during the experimentation. The instrumental variable is the interaction term between the geographic constraint on experimentation site i's land supply (determined by its land slope) with the temporal variations in the national interest rate in year t. y_p is the indicator of whether policy p eventually was rolled out to the entire country; ψ_i is a full set of county fixed effects; δ_m is a full set of ministry fixed effects, and ν_t is a full set of time fixed effects.[14]

Consistent with Chen and Kung (2016), we find that the interaction between the land suitability index and temporal interest rate strongly and positively predicts the land revenue received by the local government in a specific year (see Appendix Table A.9 for the first stage results). Table 5, Panel A presents the second stage results. We find robust positive coefficients of instrumented land revenue at experimentation sites on the corresponding policy's national roll-out. The estimates imply that, if an experimentation-participating county's land revenue doubles in a given year due to an exogenous windfall, the policy at trial will be 2.9 percentage points more likely to eventually roll out to the entire country.

In other words, when policy experimentation is conducted in localities experiencing temporal shocks that could improve the policy outcome, the central government does not fully discount these factors, but instead mistakenly attributes them (at least partially) to the underlying policy effectiveness, resulting in biased policy learning and policy choices.

Policy learning and politician-specific shocks When evaluating experimentation outcomes, does the central government exclude politician-specific shocks that are orthogonal to the underlying policy effectiveness? In particular, we examine whether an increase in local politicians' career incentives (and thus increased effort as shown in Section 5) due to political turnover affects the central government's policy learning and increases the likelihood that the policy at trial be evaluated as successful.

We focus on local politicians' turnover taking place *after* the beginning of policy experimentation in the local region, and we distinguish whether the turnover leads to an increase or decrease in local politicians' career incentives as measured in Section 5.2. This allows us to isolate changes in local politicians' career incentives that are unrelated to either the underlying effectiveness of the policy at trial, or the local government's initial participation in the experiment.

14. Following Chen and Kung (2016), we also control for characteristics at the county level (log population and local GDP growth rate), at politician level (their age, educational attainment, whether they are a member of the Youth League, previous prefectural government experience, birth-county connection with the prefectural leader, and current year in office).

Specifically, we estimate the following econometric model:

$$y_p = \alpha \cdot Turnover_{ip} + \beta \cdot Turnover_{ip} \times \Delta Incentive_{ip} + \gamma_t + \delta_m + \theta_n + \varepsilon_{ipmnt},$$

where y_p is the indicator of experiment p being evaluated as successful and rolling out to the entire country; $Turnover_{ip}$ is the indicator of a change in the party secretary of the prefecture i during the experimentation period of policy p; $\Delta Incentive_{ip}$ is the difference in career incentives between the incumbent at the beginning of the experiment and that of his or her immediate successor, following the calculation described in Appendix B.1; γ_t is a full set of year fixed effects; δ_m is a full set of ministry fixed effects; and θ_n is a full set of province fixed effects.

Table 5, Panel B presents the results. We observe a consistent pattern that, for experiments that are implemented in localities that experienced more local political turnover *after* the start of the experiment, the corresponding policy at trial is substantially more likely to be evaluated as successful and become national policy. This is especially the case if the local political turnover results in an increase in local politicians' career incentives relative to the outgoing politicians. According to our estimates, if an experimentation-participating prefecture experiences a political rotation that increases the local politician's career incentive by 1 standard deviation, then the probability that the policy at trial eventually rolls out to the whole country would increase by 7.1 percentage points. This suggests that, when policy experiments are conducted in localities that experience politician-related shocks that could improve the policy outcome, the central government attributes the outcome at least partially to policy effectiveness, again resulting in biases in policy learning. We do not observe similar effects with the rotation of politicians that preceded the policy experiments, suggesting that there is no generic pattern on increased roll-out associated with political rotation *per se* (see Appendix Table A.10).

Positively selected experimentation and national roll-out Finally, we document that policy experiments closer to being representative in their implementation are *less* likely to roll out to become nationwide policies. Appendix Table A.11 presents results where we predict the likelihood of policy roll-out based on the underlying t-statistics from the representativeness tests of the corresponding experiment's site selection. We observe a robust pattern that deviations from representativeness are associated with *higher* chances of rolling out to the entire country. Intriguingly, this association is much stronger among experiments that are ex-ante uncertain. While only suggestive, these results are consistent with the interpretation that the central government does not fully take into account the fact that positively selected experiments are less likely to reveal sub-optimal policies.

6.2 National policy outcomes

Information loss due to unrepresentative sample Experimentation reflects the central government's desire to learn about the mapping from policy to outcome. A large degree of heterogeneity in the mapping from policy to outcome would imply a certain degree of information loss due to deviation from representativeness.

We illustrate such information loss using the context of a specific policy experiment on local fiscal empowerment. In order to foster economic growth across Chinese counties, the central government initiated an experiment that allows provincial governments to bypass prefectural governments and directly administer the counties within their jurisdiction, effectively providing fiscal empowerment to the counties participating in the experiment. Between 2003 and 2013, more than 1,100 counties were selected as experimentation sites (see Appendix A.2 for details).

We observe that the experimentation sites were positively selected during the first half of experiment (t-stats of a t-test on pre-experimentation GDP per capita between experimentation and non-experimentation counties are as high as 9.110 during the first two years of experimentation), then shifted toward negative selection (t-stats decrease to -6.156 by 2007) and moved closer to representative selection toward the end of the experiment (see Appendix Figure A.17).

We find considerable heterogeneity in the effects of such experimentation on local economic development. Using a staggered event study design to estimate the treatment effects on local economic performance among experimentation counties in the early rounds (positively selected) and the later rounds (negatively selected), while controlling for county and year fixed effects, counties that have higher pre-experimentation GDP per capita benefited from the experiment, while the poorer counties experienced worse subsequent local economic development (see Appendix Figure A.18).[15]

In fact, the fiscal empowerment experiment, had it been rolled out to the entire country, would generate a net zero effect with both winners and losers (see Appendix Figure A.20 for the distribution of the projected treatment effects of the local fiscal empowerment for each county in China). This case of the local fiscal empowerment experiment demonstrates that unrepresentativeness of experimentation sites disproportionately highlights the positive effects of the policy, which could mask the unequal nature of the policy and in turn bias the central government's policy choices.

15. Such patterns of heterogeneity by pre-experimentation local economic conditions do *not* merely reflect a general equilibrium effect or an early-mover advantage of the local fiscal empowerment scheme. Less-developed counties participating in the experiment during the early rounds also experienced a negative policy treatment effect in magnitudes similar to the less-developed experimentation sites in later rounds (see Appendix Figure A.19).

Consequences due to positive selection of experimentation sites We next examine the overall effects of national policies originating from non-representative experimentation. When such policies roll out to the entire country, do localities similar to experimentation sites benefit more from the new policy?

For each experiment that eventually leads to national policies, we calculate the Mahalanobis distance between localities that participated in the experiment and those that did not (M_{cp}). The distance is calculated based on a vector of pre-experimentation local conditions: local GDP per capita, local fiscal income, and fiscal expenditure. We then examine, among localities that did not participate in the experimentation, whether a national policy leads to faster local economic growth when a specific county is socioeconomically similar to the experimentation sites of that corresponding policy. We estimate the following specification:

$$Growth_{cpt} = \alpha \cdot M_{cp} + \gamma_c + \sigma_t + \eta_p + \epsilon_{cpt}, \tag{5}$$

where $Growth_{cp}$ is (non-experimentation) county c's GDP growth after policy p rolls out to the entire country, γ_c is a full set of county fixed effects, σ_t is a full set of year fixed effects, and η_p is a full set of policy fixed effects.

The results are presented in Table 6, Panel A. We observe that localities that did not participate in an experiment but are socioeconomically similar to the experimentation sites benefit significantly *more* from the policies when they roll out to the rest of the country. This result is robust to different indices chosen to compute the distance (See Appendix Table A.12).

These results suggest that policies originating from unrepresentative experiments differentially benefit some regions over others depending on the sample composition of the experimentation sites. Given that the experimentation sites are overwhelmingly positively selected in terms of local socioeconomic conditions, this would generate distributional consequences: positive selection of sites may produce a portfolio of policies that systematically favor the more developed regions at the expense of their less-developed counterparts, thus leading to greater inter-regional inequality throughout China.

Consequences due to endogenous efforts during experimentation Moving to factors related to endogenous efforts, we next investigate the effects of national policies originating from experiments that were implemented by local politicians with strong career incentives: when these policies roll out to the entire country, do local governments whose officials have levels of career incentives similar to the experimentation sites benefit more from the new policy?

We follow an empirical approach similar to the previous sub-section: for policy exper-

iments that eventually lead to national policies, we calculate the Mahalanobis distance on the local government career incentives between localities that participated in the experiment (when the experiments started) and those that did not (when the policies rolled out to the entire country). We then estimate, among localities that did not participate in the experiment, whether national policies lead to faster local economic growth when a specific county is similar to the experimentation sites of the corresponding policy in terms of local government career incentives.

The results are presented in Table 6, Panel B. We find that non-experimentation sites with local politicians facing similar career incentives as the experimentation sites are better off when trial policies roll out to the entire country. This suggests that experimentation may structurally allow for better tailoring of policies to benefit from greater politician efforts. Note that, while we use identical measures of politicians' career incentives (as described in Appendix B.1) during and after the experiments, such career incentives could be associated with a greater degrees of effort during a trial (when local efforts are showcased) than during national implementation.

Adjusting for policy effects from experimentation To gauge the overall magnitude of policy experimentation's exaggerated signals and to guide adjustment on future policy learning, we estimate a "deflating coefficient" that maps policy effects observed during experimentation to effects among non-experimentation sites during national roll-out. Specifically, we first estimate the unconditional correlation between average effects of experimentation on local economic growth across experimentation sites and the average effects on non-experimentation sites when the corresponding policy rolls out to the entire country. Figure 4 presents the coefficient estimates. We find an unconditional deflating coefficient of 74%, namely, the policy effects (on local economic growth) decrease by 74% once they roll out beyond the experimentation stage. This could capture both unobservable differences of policy implementations during and out of experimentation, as well as observable differences in experimentation sites' sample selection and endogenous efforts during the experimentation. When we take into account the sample selection and career incentives of local politicians during experimentation, the deflating coefficient further increases to 85%, suggesting that the central government could improve the inference adjustment with these observable characteristics.

Complementarity between positive selection and endogenous efforts Career incentives of local politicians play an important role in explaining the positive selection of experimentation sites (as we have shown in Section 4.4); such career incentives also induce

greater exertion of effort during experimentation (as shown in Section 5). This implies that one cannot easily decompose the effects on national outcomes into positive site selection versus endogenous local efforts.[16]

Rather, there exists complementarity between positive selection and endogenous efforts. Richer localities participating in experiments are also more likely to have local politicians with higher career incentives and thus will exert greater efforts during an experiment. On the contrary, non-experimentation sites are more likely to be localities where socioeconomic development is less advanced, and local politicians face weaker career incentives. Therefore, the negative selection of the non-experimentation sites cannot be compensated by greater efforts exerted by local politicians. In fact, the negative selection would be compounded by the additional disadvantage of the lack of local political incentives during policy implementation.

7 Conclusion

In this project, we systematically examine China's policy experimentation over the past four decades, one of the largest undertakings of systematic policy learning in recent history. We present three sets of results. First, policy experimentation sites are substantially positively selected, and misaligned incentives across political hierarchies account for much of the observed positive selection. Second, experimental situation during policy experimentation is unrepresentative: local politicians exert strategic efforts and allocate more resources during experimentation that may exaggerate policy effectiveness. Third, the positive sample selection and unrepresentative experimental situation are not fully accounted for when the central government evaluates experimentation outcomes, which would bias policy learning and national policies originated from the experiments.

We highlight that policy learning and policy experimentation inevitably take place in complex environments with various constraints and distortions. The political and bureaucratic environment could affect the initiation of policy experimentation, its structure and implementation, and the bias in the information one may gather from an experiment. Our findings stand in contrast with theoretical work analyzing experimentation in federalist environments featuring voluntary local initiatives (Mukand and Rodrik 2005; Callander and Harstad 2015; Myerson 2015).[17] While misaligned incentives between the central and local governments generate sub-optimal learning, rather than the informational free-

16. In fact, allowing local political incentives to affect site selection may be an important mechanism through which the central government induces politicians' efforts during experimentation.

17. Cheng and Li (2019) notes, however, that the uncertainty related to citizens' inference on politicians' types could induce politicians to over-experiment even in a decentralized environment.

riding and under-experimentation observed in federalist systems, political centralization in a context such as China, where local governments compete and differentiate in order to increase their chances of promotion, could induce over-experimentation.

Our examination of China's policy experiments suggests that, while experimentation can facilitate reform and prevent policy disasters, one needs to pay attention to the manner in which policy experiments are conducted, as more information does not necessarily result in better decision making.[18] Our findings that policies originating from unrepresentative experimentation could disproportionately benefit richer regions demonstrate yet another manifestation of regulatory capture — by systematically biasing the information that decision makers gather during the policy learning process — in addition to pure regulatory capture (e.g., Stigler 1971), capture through corruption (e.g., Shleifer 1996), and capture through enforcement (e.g., Glaeser and Shleifer 2003), recent literature has documented more subtle forms of cognitive capture of regulators (e.g., Johnson and Kwak 2011) and capture through philanthropic giving and strategic advocacy (Bertrand et al. 2020).[19] Moreover, our findings point to a fundamental trade-off that the central government faces: structuring political incentives in order to stimulate politicians' effort to improve policy outcomes, while making sure that such incentives are not exaggerated during the experimentation phase, so that policy learning remains unbiased. Future work on mechanism design solutions that could improve the efficiency of policy learning could be of great policy relevance and importance.

18. It is important to note that our work does *not* address the overall benefits of experimentation (as opposed to implementing national policies without going through any experimentation). This is an important avenue for future work.

19. Our evidence of informational capture through politically connected government officials also relates to the growing body of work documenting the costs and distortions associated with political patronage, specifically in China's context (e.g., Fisman and Wang 2015; Fisman et al. 2020).

References

Acemoglu, Daron, David Y Yang, and Jie Zhou. 2021. "Political Pressure and the Direction of Research: Evidence from Chinas Academia." *Working paper.*

Aghion, Philippe, Patrick Bolton, Christopher Harris, and Bruno Jullien. 1991. "Optimal learning by experimentation." *The review of economic studies* 58 (4): 621–654.

Allcott, Hunt. 2015. "Site selection bias in program evaluation." *The Quarterly Journal of Economics* 130 (3): 1117–1165.

Banerjee, Abhijit V, Sylvain Chassang, Sergio Montero, and Erik Snowberg. 2020. "A theory of experimenters: Robustness, randomization, and balance." *American Economic Review* 110 (4): 1206–30.

Bergquist, Lauren, Benjamin Faber, Thibault Fally, Matthias Hoelzlein, Edward Miguel, and Andres Rodriguez-Clare. 2019. "Scaling Agricultural Policy Interventions: Theory and Evidence from Uganda." *Unpublished manuscript, University of California at berkeley.*

Bertrand, Marianne, Matilde Bombardini, Raymond Fisman, Brad Hackinen, and Francesco Trebbi. 2020. *Hall of Mirrors: Corporate Philanthropy and Strategic Advocacy.* Technical report. Boston University-Department of Economics.

Cai, Hongbin, Daniel Treisman, et al. 2009. "Political decentralization and policy experimentation." *Quarterly Journal of Political Science* 4 (1): 35–58.

Callander, Steven. 2011. "Searching for good policies." *American Political Science Review,* 643–662.

Callander, Steven, and Bård Harstad. 2015. "Experimentation in federal systems." *The Quarterly Journal of Economics* 130 (2): 951–1002.

Cao, Yuanzheng, Yingyi Qian, and Barry R Weingast. 1999. "From federalism, Chinese style to privatization, Chinese style." *Economics of Transition* 7 (1): 103–131.

Chandler, Alfred Dupont. 1962. *Strategy and structure: Chapters in the history of the industrial enterprise.* Vol. 120. MIT press.

Chen, Ting, and JK-S Kung. 2016. "Do land revenue windfalls create a political resource curse? Evidence from China." *Journal of Development Economics* 123:86–106.

Cheng, Chen, and Christopher Li. 2019. "Laboratories of democracy: Policy experimentation under decentralization." *American Economic Journal: Microeconomics* 11 (3): 125–54.

Davis, Jonathan M.V., Jonathan Guryan, Kelly Hallberg, and Jens Ludwig. 2017. "The Economics of Scale-up." *NBER Working Paper.*

DellaVigna, Stefano, and Woojin Kim. 2021. "Policy Diffusion and Polarization across U.S. States."

DellaVigna, Stefano, and Elizabeth Linos. 2020. *Rcts to scale: Comprehensive evidence from two nudge units.* Technical report. National Bureau of Economic Research.

Dewatripont, Mathias, and Gerard Roland. 1995. "The design of reform packages under uncertainty." *The American Economic Review,* 1207–1223.

Fang, Hanming, Chang Liu, and Li-An Zhou. 2020. *Window Dressing in the Public Sector: A Case Study of Chinas Compulsory Education Promotion Program.* Technical report. National Bureau of Economic Research.

Fisman, Raymond, Jing Shi, Yongxiang Wang, and Weixing Wu. 2020. "Social ties and the selection of China's political elite." *American Economic Review* 110 (6): 1752–81.

Fisman, Raymond, and Yongxiang Wang. 2015. "The mortality cost of political connections." *The Review of Economic Studies* 82 (4): 1346–1382.

Gechter, Michael, and Rachael Meager. 2021. "Combining Experimental and Observational Studies in Meta-Analysis: A Mutual Debiasing Approach."

Glaeser, Edward L, and Andrei Shleifer. 2003. "The rise of the regulatory state." *Journal of economic literature* 41 (2): 401–425.

Han, Li, and James Kai-Sing Kung. 2015. "Fiscal incentives and policy choices of local governments: Evidence from China." *Journal of Development Economics* 116:89–104.

Hayek, Friedrich August. 1978. *Law, legislation and liberty, volume 1: Rules and order.* Vol. 1. University of Chicago Press.

Heilmann, Sebastian. 2008a. "From local experiments to national policy: the origins of China's distinctive policy process." *The China Journal,* no. 59, 1–30.

———. 2008b. "Policy experimentation in Chinas economic rise." *Studies in Comparative International Development* 43 (1): 1–26.

Hirsch, Alexander V. 2016. "Experimentation and persuasion in political organizations." *American Political Science Review* 110 (01): 68–84.

Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini. 2019. *How research affects policy: Experimental evidence from 2,150 brazilian municipalities.* Technical report. National Bureau of Economic Research.

Huang, Xiulan. 2000. "On Policy Experimentations in the Reform and Open Up Process." *Probe* 3:66–69.

Johnson, Simon, and James Kwak. 2011. *13 bankers: The Wall Street takeover and the next financial meltdown.* Vintage.

Mehmood, Sultan, Shaheen Naseer, and Daniel L Chen. 2021. *Training Policymakers in Econometrics.* Technical report. Working Paper.

Montinola, Gabriella, Yingyi Qian, and Barry R Weingast. 1995. "Federalism, Chinese style: the political basis for economic success in China." *World politics,* 50–81.

Mukand, Sharun W, and Dani Rodrik. 2005. "In search of the holy grail: policy convergence, experimentation, and economic performance." *American Economic Review* 95 (1): 374–383.

Myerson, Roger. 2015. "Local Agency Costs of Political Centralization." *U. Chicago Working Paper.*

Naughton, Barry. 1996. *Growing out of the plan: Chinese economic reform, 1978-1993.* Cambridge university press.

North, Douglass C, et al. 1990. *Institutions, institutional change and economic performance.* Cambridge university press.

Qian, Yingyi. 2002. "How reform worked in China."

Qian, Yingyi, Gerard Roland, and Chenggang Xu. 2006. "Coordination and experimentation in M-form and U-form organizations." *Journal of Political Economy* 114 (2): 366–402.

Rawski, Thomas G. 1995. "Implications of China's reform experience." *China Q.*, 1150.

Rogger, Daniel, and Ravi Somani. 2018. *Hierarchy and information.* The World Bank.

Roland, Gerard. 2000. *Transition and economics: Politics, markets, and firms.* MIT press.

Sachs, Jeffrey D. 2006. *The end of poverty: Economic possibilities for our time.* Penguin.

Shipan, Charles R, and Craig Volden. 2006. "Bottom-up federalism: The diffusion of antismoking policies from US cities to states." *American journal of political science* 50 (4): 825–843.

Shleifer, Andrei. 1996. "Origins of bad policies: Control, corruption and confusion." *Rivista di Politica Economica.*

Snowberg, Erik, and Leeat Yariv. 2018. "Testing the waters: Behavior across subject pools." *NBER Working Paper No* 24781.

Stigler, George J. 1971. "The theory of economic regulation." *The Bell journal of economics and management science*, 3–21.

Al-Ubaydli, Omar, John A. List, and Dana Suskind. 2019. "The Science of Using Science: Towards an Understanding of the Threats to Scaling Experiments." *NBER Working Paper.*

Vivalt, Eva. 2020. "How much can we generalize from impact evaluations?" *Journal of the European Economic Association* 18 (6): 3045–3089.

Vivalt, Eva, and Aidan Coville. 2019. *How do policymakers update?*

Wang, Zhi, Qinghua Zhang, and Li-An Zhou. 2020. "Career incentives of city leaders and urban spatial expansion in China." *Review of Economics and Statistics* 102 (5): 897–911.

Williamson, Oliver E. 1975. "Markets and hierarchies: analysis and antitrust implications: a study in the economics of internal organization." *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship.*

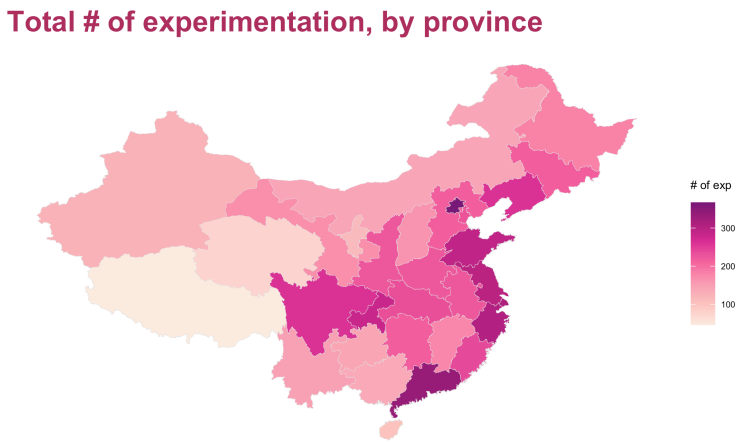
Wu, Youxi. 1995. "Analysis of Policy Experiment Methods." *Reform of Economic System* 6.

Xie, Yinxi, and Yang Xie. 2017. "Machiavellian experimentation." *Journal of Comparative Economics* 45 (4): 685–711.

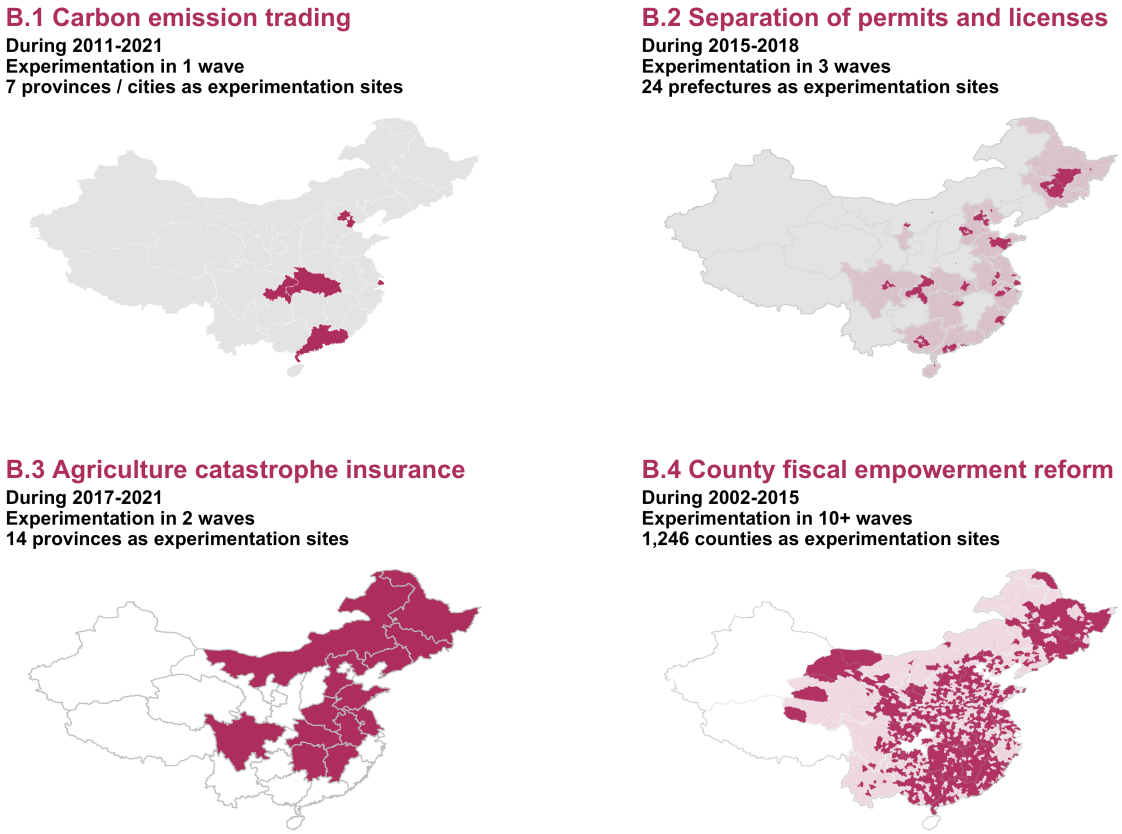
Xu, Chenggang. 2011. "The fundamental institutions of China's reforms and development." *Journal of economic literature* 49 (4): 1076–1151.

Zhou, Wang. 2013. *Study on China's Experimental Points.* Tianjin People's Press.

Figures and tables



Panel A: Spatial distribution of policy experimentations



Panel B: Examples of policy experimentation

Figure 1: These maps plot the spatial distribution of policy experimentation in China. Panel A counts the total number of policy experiments that each province has been involved in (including experiments at prefectural and county levels). Panels B.1 and B.2 show two policies that eventually rolled out to the entire country. The regions shaded in grey indicate parts of the country that eventually received the policy. Panels B.3 and B.4 show two policies that did not eventually roll out. The experimentation sites are marked in red, and the corresponding provinces are marked in pink.

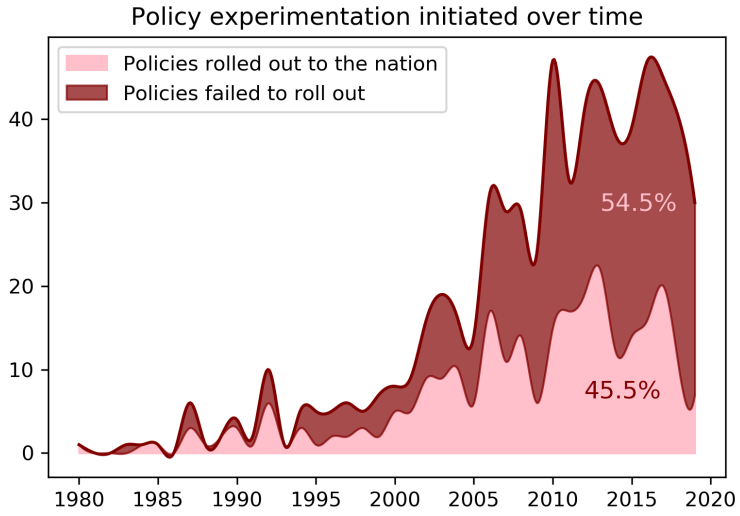
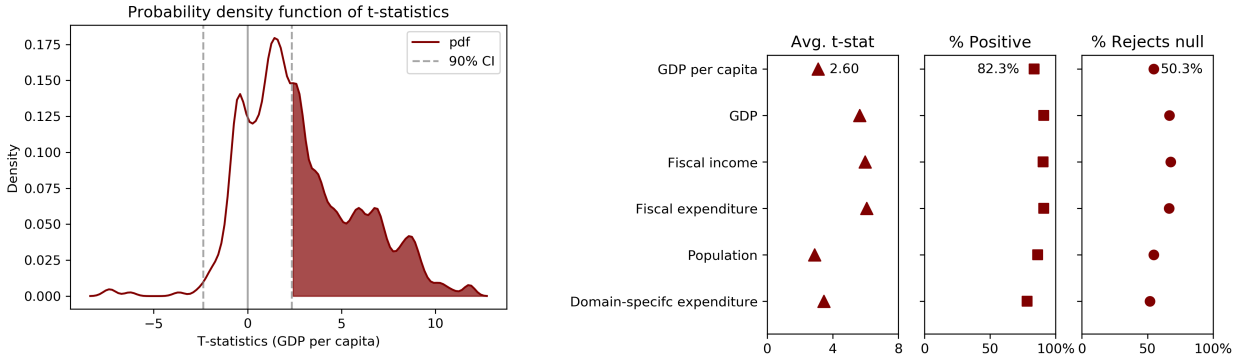


Figure 2: This figure plots the number of policy experiments initiated over time. The share of successful experiments that eventually rolled out to the entire country is indicated by the area shaded in pink; the share of unsuccessful policies that failed to roll out to the entire country is indicated by the area shaded in red.



Panel A Panel B

Figure 3: This figure shows descriptive facts on the representative test. Panel A plots the t-statistics distribution from the representativeness test, calculated based on GDP per capita, to serve as an example. Panel B extends the list to more socioeconomic characteristics, and reported the mean of t-statistics, the percentage of policies with t-stat> 0, and the percentage of tests where we can reject the null hypothesis $H_0 : \bar{Y}(0) = \bar{Y}(1)$ in three sub-panels, respectively. To calculate the t-statistics, we compare the average pre-experimentation characteristics between those jurisdictions chosen as experimentation sites, and their peers at the same hierarchical level that were not chosen as experimentation sites within each test. The grey vertical lines in panel A represent the average critical value at 90% confidence level among all t-tests.

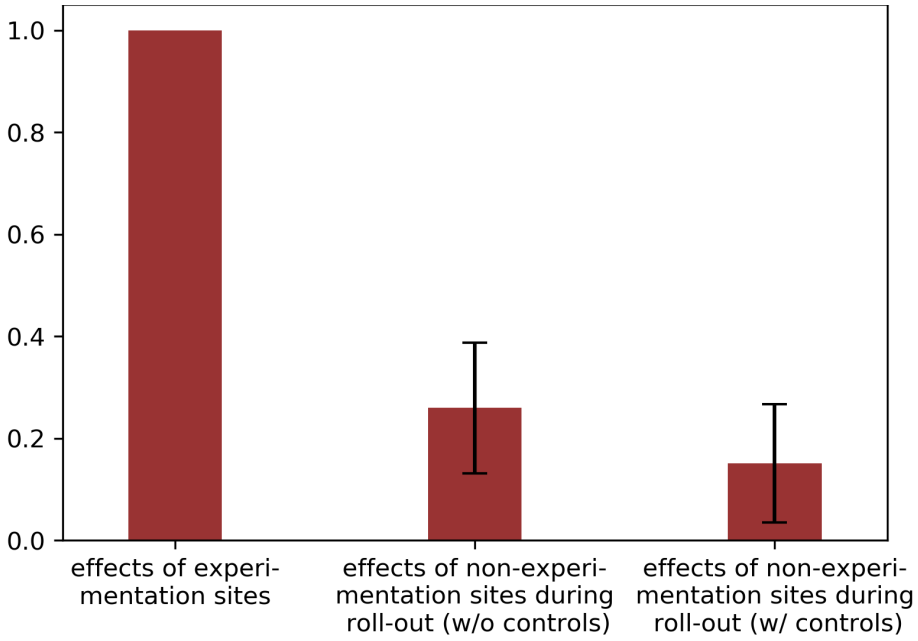


Figure 4: The figure shows the difference between the average treatment effects of experimentation sites (standardized to 1), and the average treatment effects of non-experimentation sites during policy roll-out. The whiskers illustrate the 95% confidence intervals of the point estimate for this deflator. Treatment effects are measured by the growth rate of GDP per capita, in logarithm terms. Standard errors are clustered at policy level.

Table 1: Summary statistics of policy experimentation

	# of exp.	# of rounds	# of sites	% roll-out	Avg. t-stats	% repre-sentative
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: **Full sample**						
Overall	633	2.9	19.1	44.7	2.60	49.7
National	594	2.9	19.7	46.0	2.70	48.2
×Completed	494	3.0	18.4	54.5	2.80	45.3
×Ongoing	100	2.7	26.2	4.0	2.15	63.1
Subnational	39	2.8	9.1	25.6	1.04	74.2
×Completed	35	2.9	9.7	25.7	1.08	75.0
×Ongoing	4	1.5	4.2	25.0	0.64	66.7
Panel B: **By policy domain**						
Resource, energy & environment	80	2.5	12.1	40.0	2.20	63.2
Market supervision	77	2.5	11.4	49.4	3.22	32.8
Agriculture	57	3.4	32.9	31.6	1.89	62.5
Education	54	3.1	43.0	48.1	2.75	34.0
Finance	53	2.5	6.3	49.1	5.27	36.4
Tax & fiscal policy	41	3.3	10.7	53.7	2.85	55.9
Commerce & trade	36	4.3	17.9	41.7	3.70	20.7
Population & health	35	3.1	22.9	48.6	2.07	41.2
Domestic affairs	31	2.8	16.6	32.3	1.72	57.7
Development & reform	28	3.2	25.5	39.3	2.13	55.0
Industry & information technology	27	2.6	21.2	40.7	4.00	40.0
Labor & personnel	22	3.2	10.2	50.0	2.25	50.0
Transportation	20	2.1	9.7	60.0	0.93	84.2
Others	33	3.2	37.9	72.7	3.04	42.4
Panel C: **By ex-ante certainty**						
Certain	190	3.3	22.7	63.7	2.12	46.8
Uncertain	404	2.8	18.3	37.6	2.95	48.7
Panel D: **By complexity**						
Single-ministry	451	2.2	14.3	42.6	2.65	52.1
Multi-ministry	143	5.2	37.0	56.6	2.84	36.7
Panel E: **By sign-up process**						
Opt-in	270	3.2	30.6	45.9	3.01	36.3
Top-down	324	2.8	11.8	45.7	2.53	61.4
Panel F: **By administrative level**						
Province level	198	1.4	4.8	36.9	1.11	72.1
City level	260	3.1	10.6	59.6	4.30	28.9
County level	136	4.9	58.9	33.1	1.55	57.0

Note: This table reports the summary statistics for our policy experimentation sample. In Panel A, we present information on all 633 experiments, and disaggregate them by national experiments (594) and subnationl ones (39). In Panels B to F, we only focus on national experiments.

Table 2: Political incentives and policy experimentation

	Engage in experimentation		
	(1)	(2)	(3)
Panel A: Local politicians' career incentive			
Career incentive	1.397*	1.405*	1.309*
	(0.796)	(0.824)	(0.791)
# of obs.	7630	7630	7630
Mean of DV	1.059	1.059	1.059
Prefecture controls	No	No	Yes
Politician controls	No	Yes	Yes
Year FE	Yes	Yes	Yes
Prefecture FE	Yes	Yes	Yes
Panel B: Political patronage			
Connected to minister	0.088**	0.062*	0.063*
	(0.035)	(0.036)	(0.037)
# of obs.	42884	42884	42884
Mean of DV	0.214	0.214	0.214
Controls	No	No	Yes
Year FE	No	Yes	Yes
Ministry by province FE	Yes	Yes	Yes
Panel C: Political stability concerns			
# of protests in previous year	-0.004**	-0.002**	-0.003***
	(0.002)	(0.001)	(0.0002)
# of obs.	1519	1519	757
Mean of DV	1.135	1.135	2.043
Pre-period controls	No	No	Yes
Year FE	No	Yes	Yes
Prefecture FE	Yes	Yes	Yes

Note: In this table we investigate how various forms of political distortion affect policy experimentation. In panel A, *Connection* is the indicator of whether the current minister possesses any full-time previous work experience in a given province. *Incentive*, in panel B, is the fitted probability of a prefectural party secretary's political promotion, as detailed in Appendix section B.1. Protest data is panel C is collected from Global Database of Events, Language, and Tone (GDELT). Standard errors are clustered at the province level in Panel A; and the prefecture level in Panels B and C.

Table 3: Local fiscal expenditure during policy experimentation

	Share of fiscal expenditure on experimentation-related domains					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Fiscal input among experimentation sites						
# of experiments	0.003***	0.002***	0.002***	-0.013***	-0.002*	-0.003
	(0.001)	(0.0004)	(0.0005)	(0.003)	(0.001)	(0.002)
# $\times$ career incentive				0.043***	0.009**	0.011**
				(0.007)	(0.004)	(0.005)
Panel B: Fiscal input among non-experimentation sites during national policy roll-out						
# of rolled out policies	0.001	0.001	0.001	0.001	0.001	0.001
	(0.001)	(0.0004)	(0.001)	(0.003)	(0.001)	(0.002)
# $\times$ career incentive				-0.001	-0.0004	-0.0003
				(0.005)	(0.002)	(0.003)
# of obs.	142,116	142,116	142,116	142,116	142,116	142,116
Mean of DV	0.174	0.174	0.174	0.174	0.174	0.174
County by category FE	No	Yes	Yes	No	Yes	Yes
Year by county FE	Yes	No	Yes	Yes	No	Yes
Category by year FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table estimates the impact of a policy experiment on the fiscal expenditures of its experimentation sites. We characterize six general fiscal domains, and match each policy experiment to its most closely related domain. In panel A, we investigate whether the experimentation units re-allocated fiscal resources to the corresponding fiscal domain when a policy experiment is assigned. In panel B, we investigate whether the previously non-experimentation sites exhibited similar fiscal reallocation in the year that the policy rolled out nationally. Standard errors are clustered at county level.

Table 4: Biased policy learning from experimentation

	Similarity index			
	(1)	(2)	(3)	(4)
Career incentive	-0.052**	-0.066**	-0.066**	-0.076**
	(0.027)	(0.032)	(0.033)	(0.038)
# of obs.	1,148	1,148	1,148	1,148
Mean of DV	0.980	0.980	0.980	0.980
Politician Controls	No	No	No	Yes
Policy FE	Yes	Yes	Yes	Yes
Year FE	No	No	Yes	Yes
Prefecture FE	No	Yes	Yes	Yes

Note: In this table, we investigate how a politician's career incentive affects how his policy experimentation plan differs from that of his peers. Career incentive is measured by the fitted probability of a prefectural party secretary's political promotion, as detailed in Appendix section B.1. For the outcome variable, we conduct Latent Semantic Analysis, a canonical approach from Natural Language Processing, to measure the text similarity of government documents, which is detailed in Appendix Section F. The similarity index, taking maximum over all similarity pairs between a document and all others issued by its counterpart administrations on the same policy, aims at measuring how much a local government politician differentiates from his or her colleagues in the policy design during experimentation. We exclude all the documents from single-wave experiments and first wave documents from multi-wave experiments. We also restrict the sample to the first key document issued by each experimentation site in each wave, and drop the follow-up documents issued within the same site-wave unit. Politician controls include his or her level of education and previous central experience. Standard errors are clustered at policy level.

Table 5: Irrational decomposition during experimentation evaluation

	National roll-out		
	(1)	(2)	(3)
Panel A: Land revenue windfall			
Land revenue (instrumented)	0.020***	0.039***	0.029***
	(0.002)	(0.003)	(0.003)
# of obs.	18,464	18,464	18,464
Mean of DV	0.509	0.509	0.509
Ministry FE	No	No	Yes
Year FE	Yes	Yes	Yes
County FE	No	Yes	Yes
Panel B: Political rotation			
Rotation	0.037*	0.044***	0.043***
	(0.019)	(0.012)	(0.012)
Rotation $\times$ change in career incentive	0.230***	0.153**	0.151**
	(0.062)	(0.012)	(0.007)
# of obs.	3899	3899	3899
Mean of DV	0.328	0.328	0.328
Ministry FE	No	YES	Yes
Year FE	Yes	Yes	Yes
Province FE	No	NO	Yes

Note: In this table, we investigate whether external shocks to a policy experiment's sites and the local officials affect its likelihood of being rolled out. Panel A reports the second stage of a 2SLS regression where we use the interaction term between area of land unsuitable for agricultural use and national interest rate to instrument for the land revenue received by the local government. We report the first stage results in Appendix Table A.9. We include politician-level control variables including the mean of his or her age across the period, education, past experience in the prefectural government, previous positions as Youth League party leaders, and hometown-connection with the prefectural leaders. Panel B is an analysis focusing on political rotations that happened after the selection of experimentation sites. At the experiment-by-prefecture level, we calculate the difference in career incentives between the leaving prefectural official and his immediate successor. *Rotation* is a dummy variable indicating political turnover during the experimentation, which is defined to be the period between the start of the first round of experimentation and two years after the last round. The standard errors are clustered at the province level.

Table 6: Similarity with experimentation sites and effects of policy roll-out

	GDP per capita growth		
	(1)	(2)	(3)
Panel A: Selection of experimentation sites			
M-distance between local development	-0.007***	-0.007***	-0.006***
	(0.001)	(0.001)	(0.001)
# of obs.	77,588	77,588	77,588
Mean of DV	0.0806	0.0806	0.0806
Panel B: Endogenous efforts during experimentation			
M-distance between career incentives	-0.001***	-0.002***	-0.0001
	(0.0002)	(0.0003)	(0.0002)
# of obs.	86,221	86,221	86,221
Mean of DV	0.0930	0.0930	0.0930
Policy FE	No	No	Yes
Year FE	No	Yes	Yes
County FE	Yes	Yes	Yes

Note: This table investigates how much of a policy's (in)effectiveness at the national roll-out stage can be attributed to the site selection and endogenous effort patterns at its experimentation stage. The sample includes all non-experimentation counties in years that a former policy experiment is being rolled out as a national policy. In Panel A, we look at the Mahalanobis distance between experimentation and non-experimentation counties for a given policy experiment, in terms of their socioeconomic conditions. In Panel B, we investigate Mahalanobis distance between the experimentation and non-experimentation sites in terms of political incentives, where career incentive is measured by the fitted probability of a prefectural party secretary's political promotion, as detailed in Appendix section B.1. The estimated covariance matrix in computing a Mahalanobis distance is fitted by the observed distribution of the data. Mahalanobis distances, in both panels, are standardized to mean zero and unit variance. Standard errors are clustered at county level.

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:

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

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 im-

pact 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:

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

where y_{mt} is the total number of policy experiments initiated by ministry/commission m in year t, and D^k_{mt} 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^k_{mt} 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:

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

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:

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

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

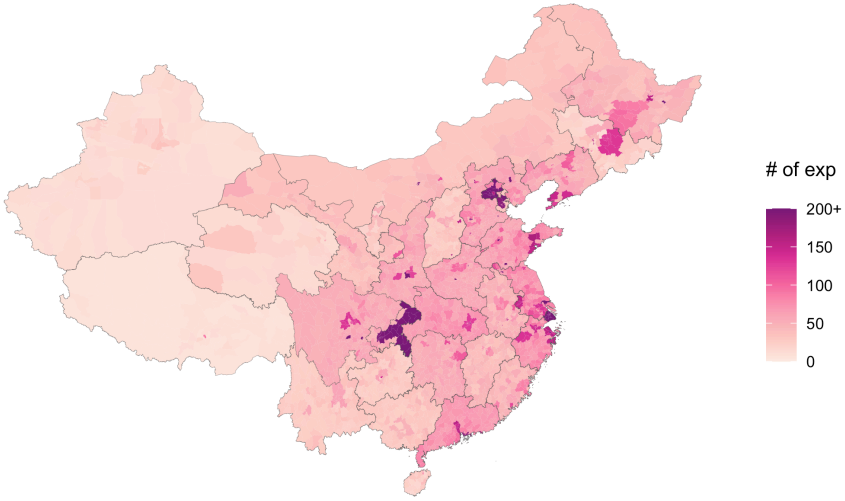
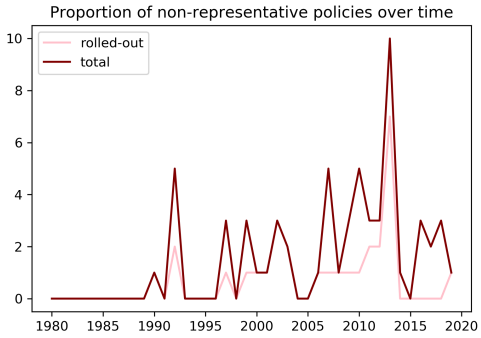
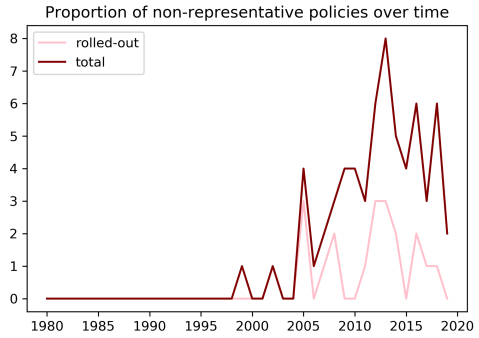


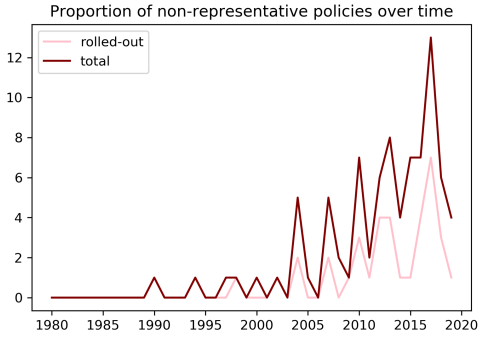
Figure A.1: This county-level map plots the spatial distribution of policy experimentation in China. A county is assigned a policy experiment if either itself or its corresponding prefecture/province serves as an experimentation site for that policy.



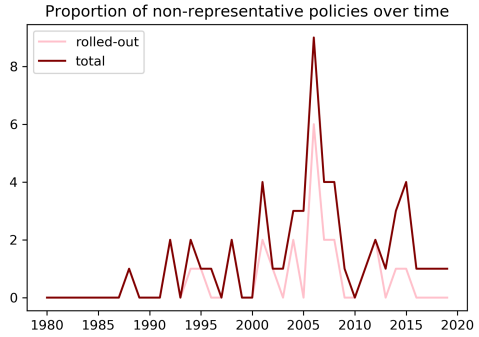
Education



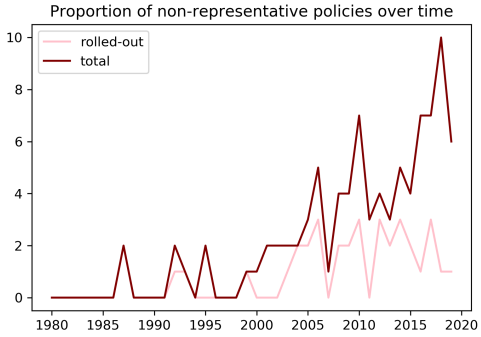
Agriculture



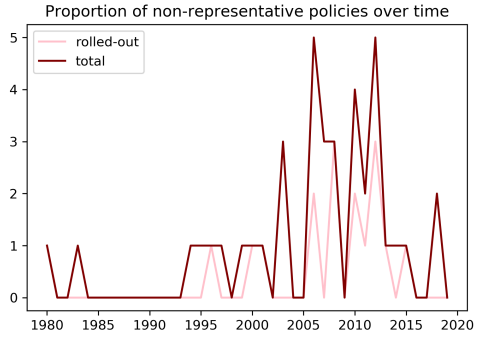
Market supervision



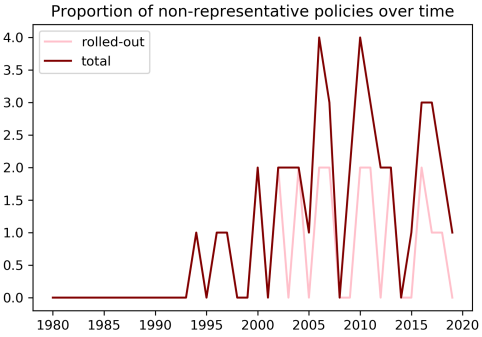
Finance & economics



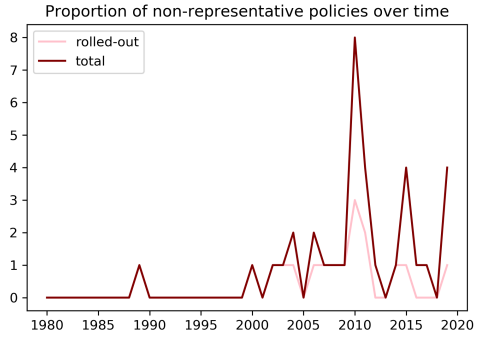
Natural resources & environment



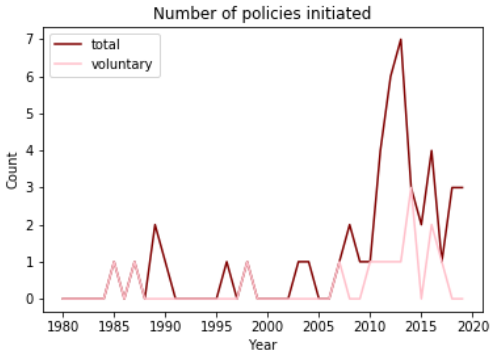
Business & commerce



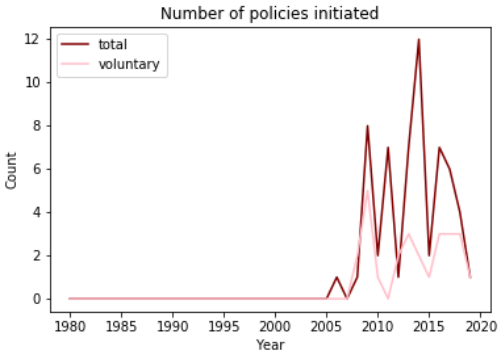
Government finance & taxation



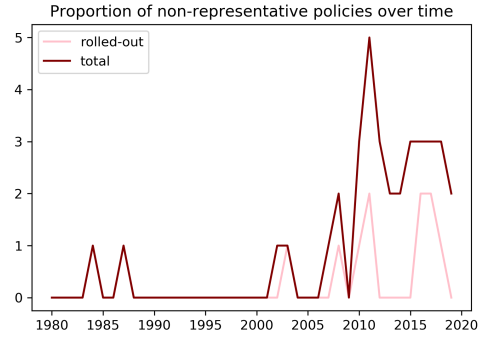
Population & health



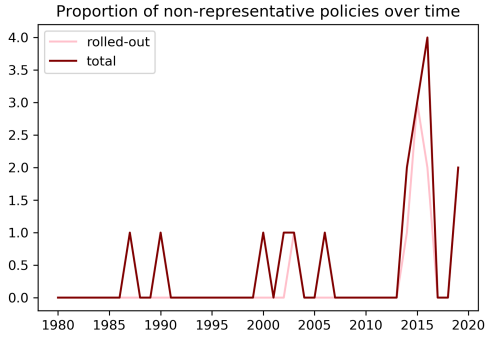
Domestic affairs



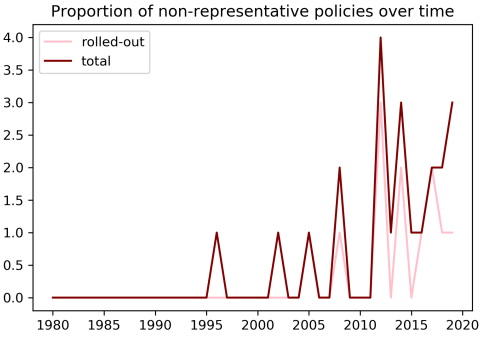
Industrial information



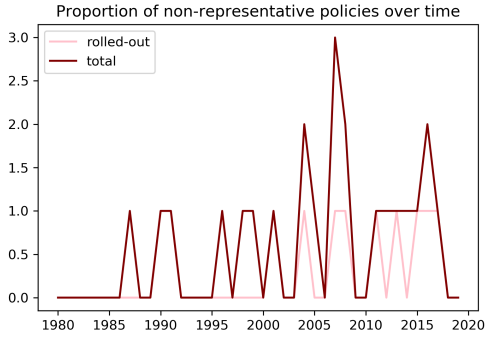
Development & reform



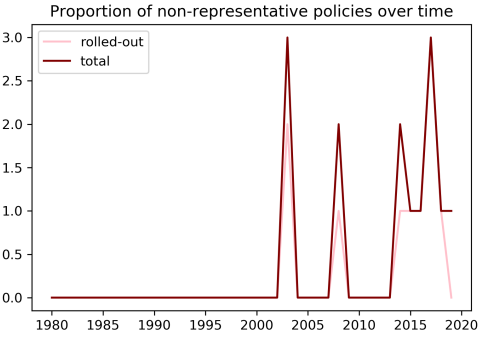
General purpose



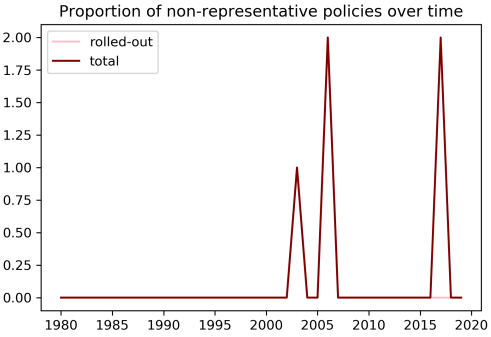
Transportation



Labor & personnel



Judiciary & supervision



Media

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

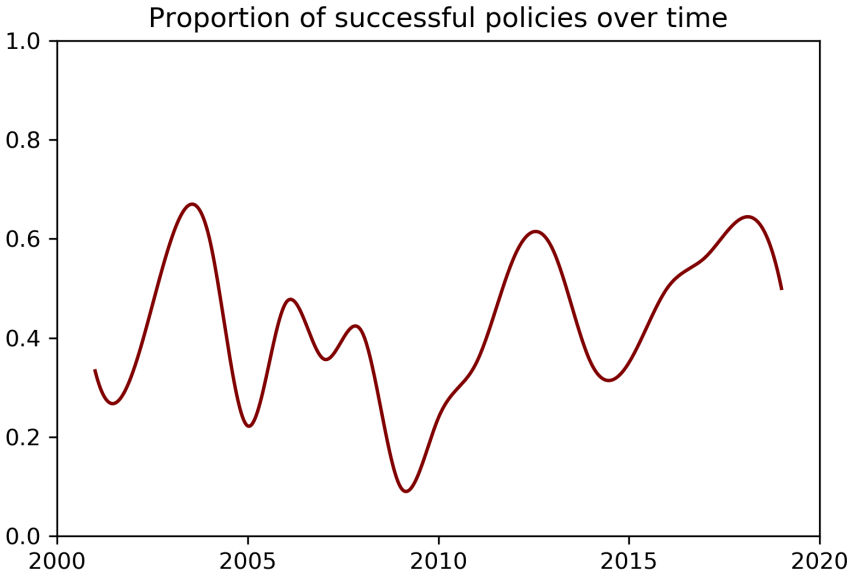


Figure A.3: This figure plots the ratio of successful policy experiments in each year. A policy experiment is defined as a “success" if it eventually rolled out to the entire nation.

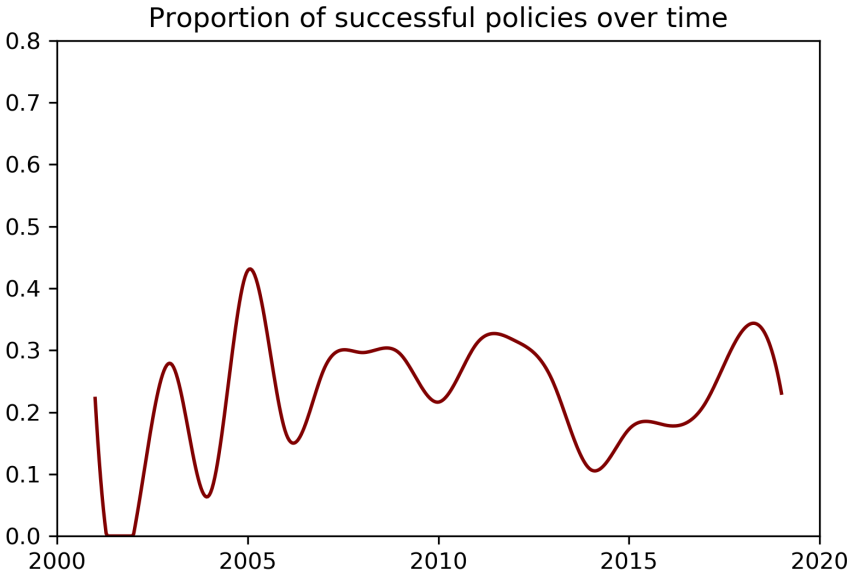


Figure A.4: Time trend of successful policy shares, 2000-2020. A policy is defined "success" if it is adopted by 2/3 of the provinces during the experimentation.

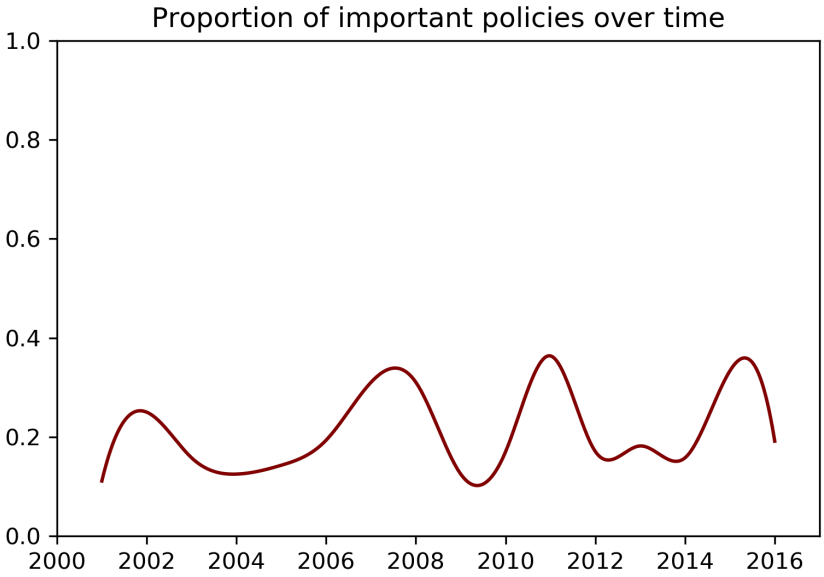


Figure A.5: Time trend of important policy shares, 2000-2017. A policy is labeled important if it is mentioned in the Five Year Plans. Those plans are both retrospective and introspective. We dropped some of the most recent years, observing that the most recent Five Year Plan is issued in 2016.

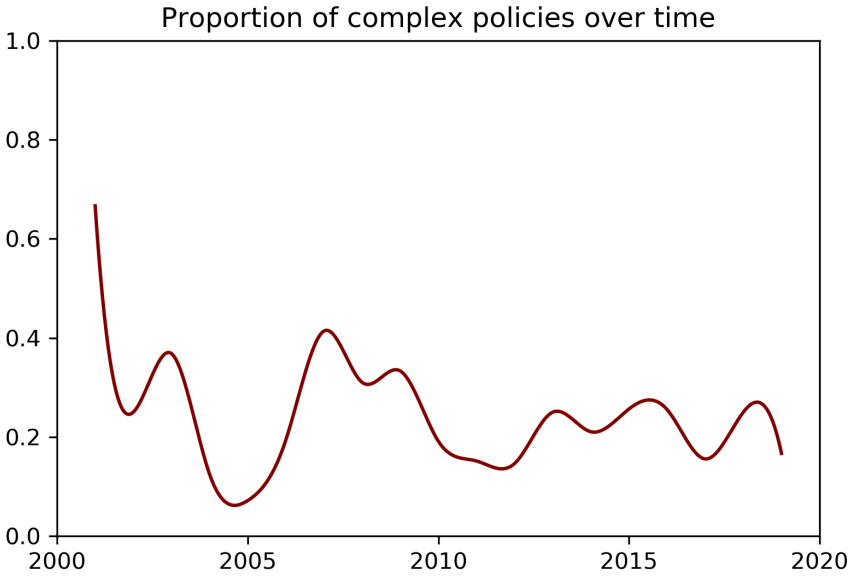


Figure A.6: This figure plots the share of policy experiments in each year that requires multi-department cooperation.

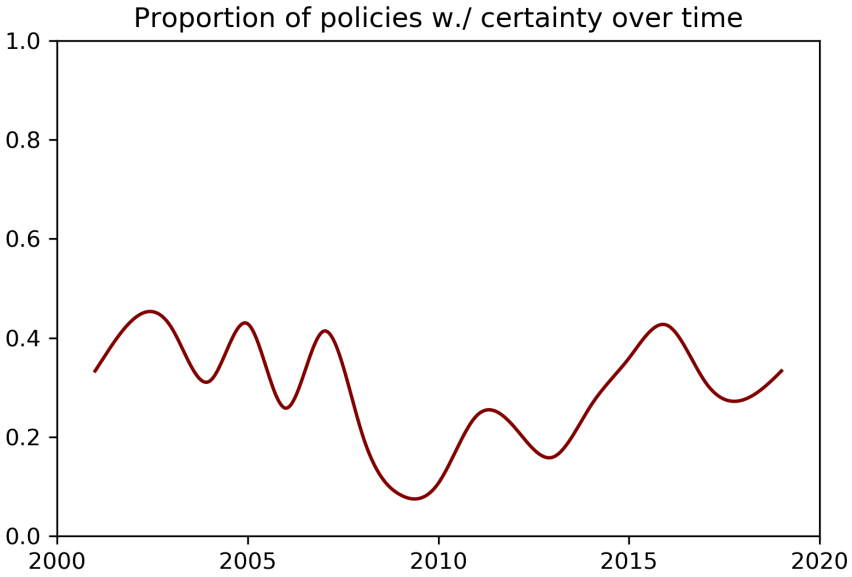


Figure A.7: This figure plots the share of policy experiments in each year that has detailed timelines of roll-out delineated in the first experimentation document.

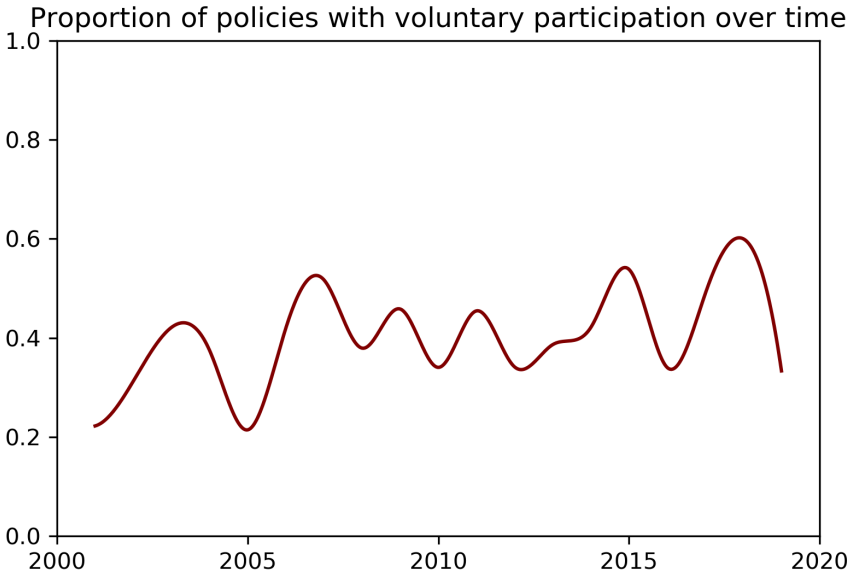


Figure A.8: This figure plots the share of policy experiments in each year that has a voluntary sign-up process for experimentation sites.

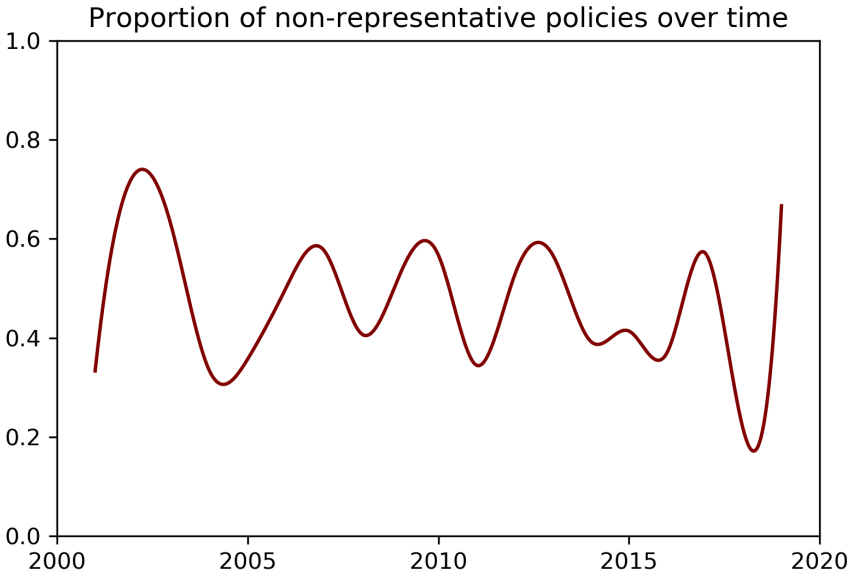
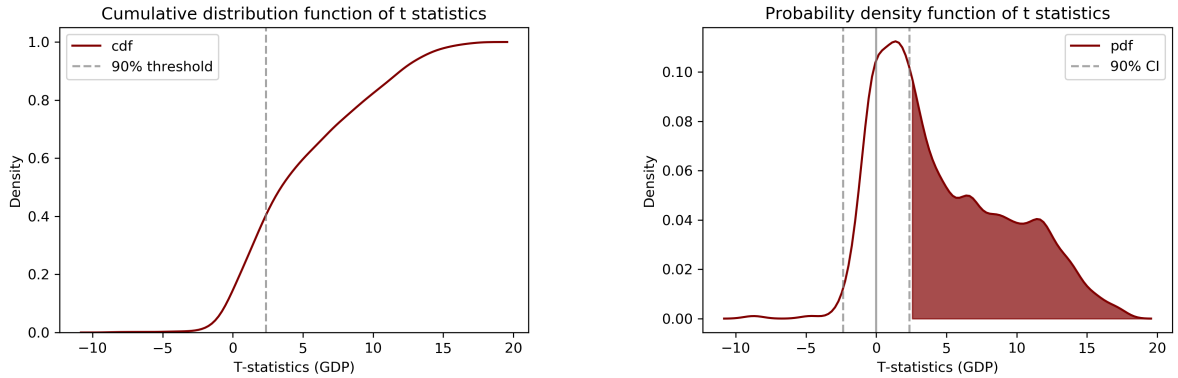
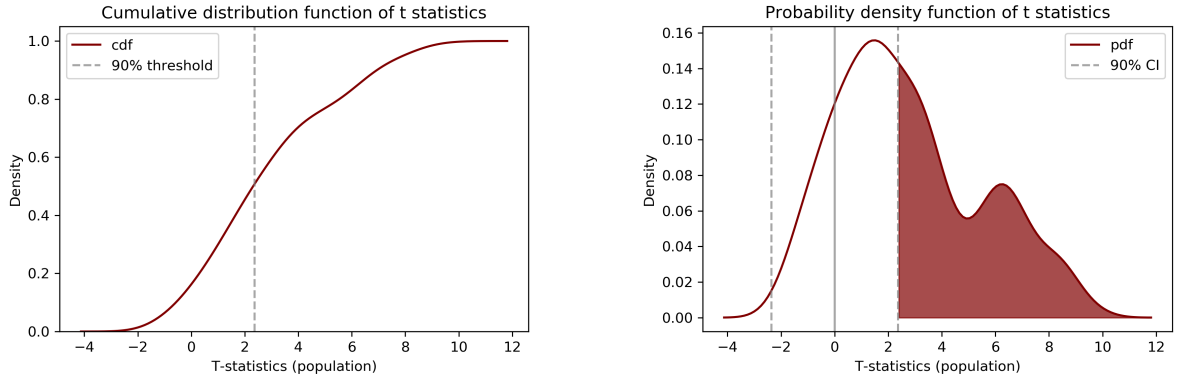


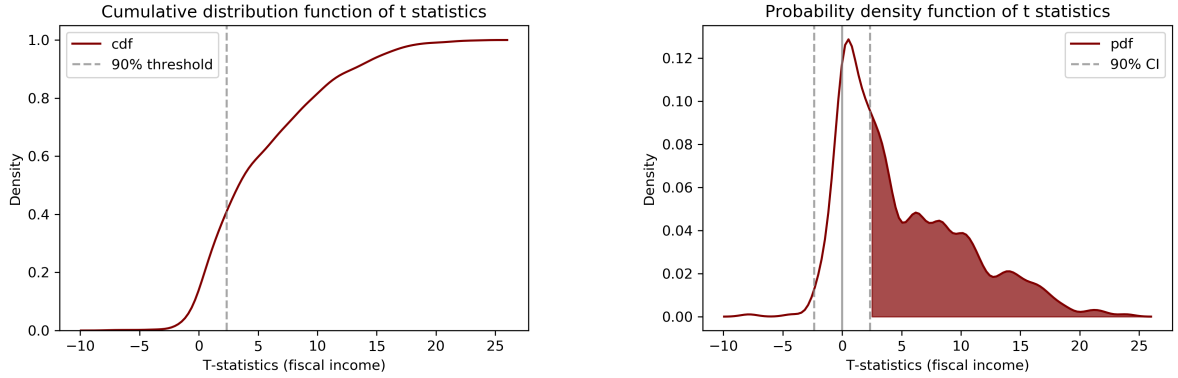
Figure A.9: This figure plots the share of non-representative policy experiments in each year. The notion non-representativeness is defined in Section 4.1.



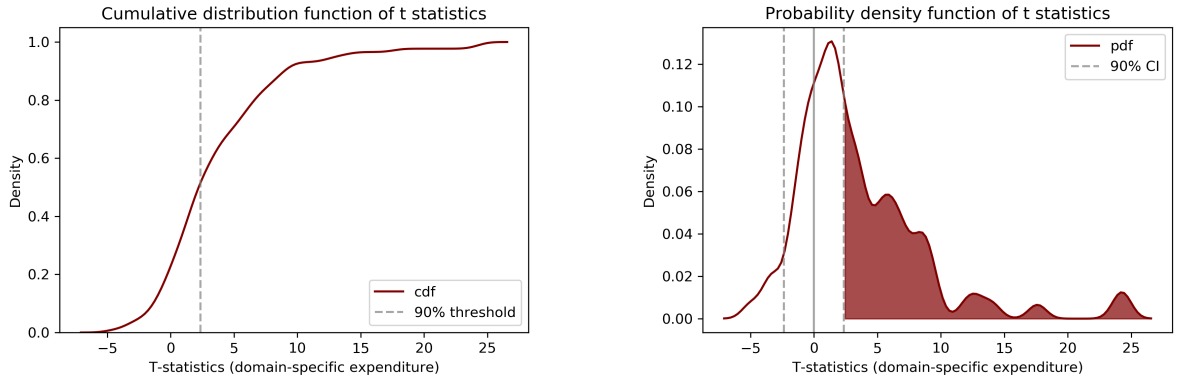
Panel A: Test with GDP



Panel B: Test with population

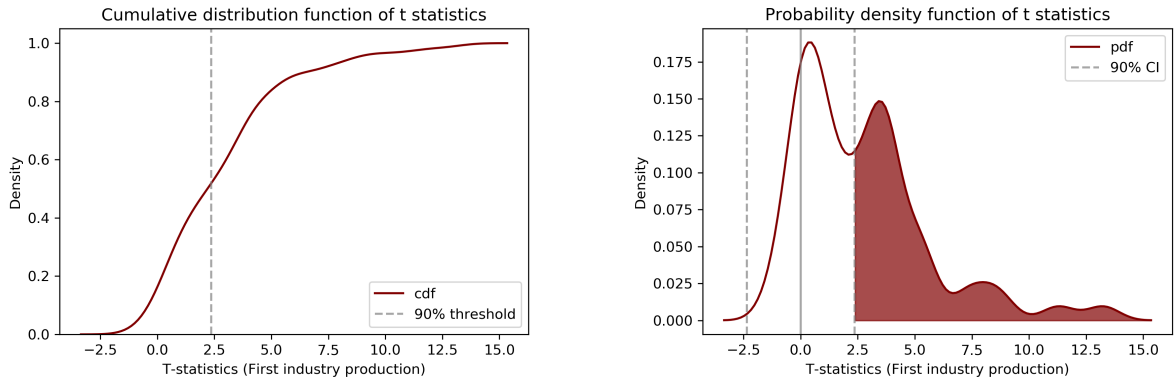


Panel C: Test with fiscal income

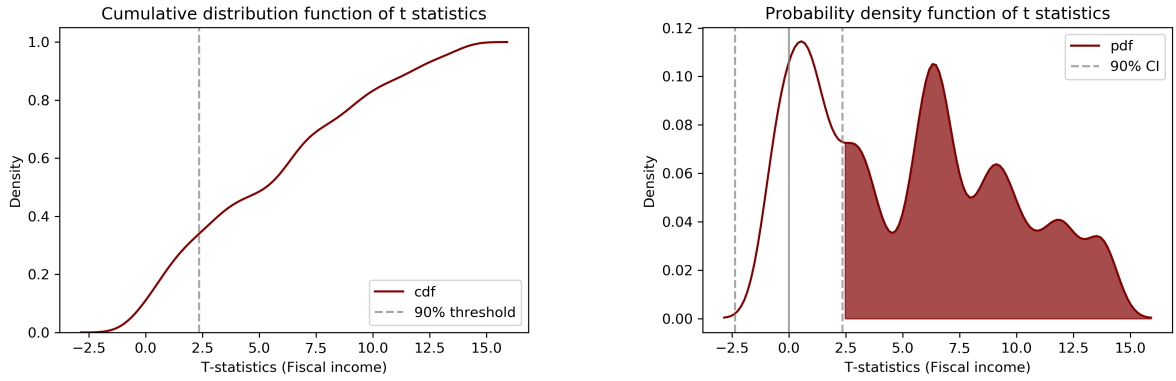


Panel D: Test with domain-specific fiscal expenditure

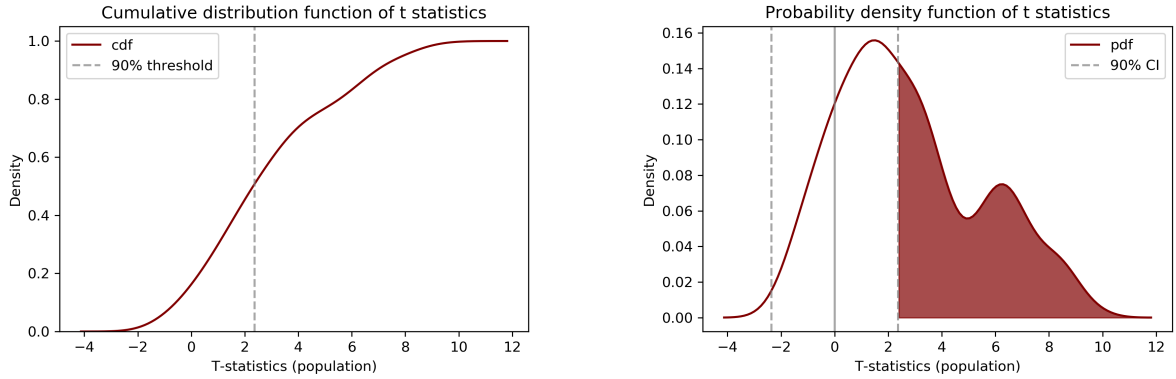
Figure A.10: This figure presents representativeness tests using alternative measures. In addition to the GDP per capita test as addressed in Figure 3, we conduct the test using total GDP in Panel A, population in Panel B, fiscal income in Panel C, and domain-specific fiscal expenditure in Panel D.



Panel A: Agricultural policies

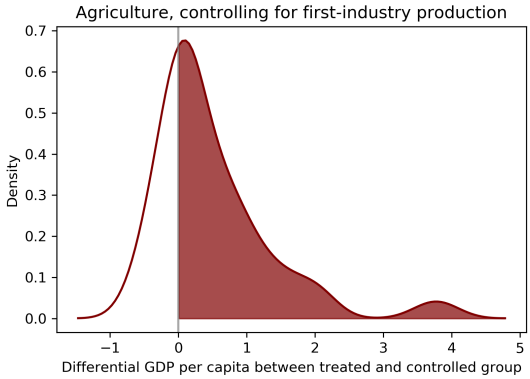


Panel B: Government finance and tax policies

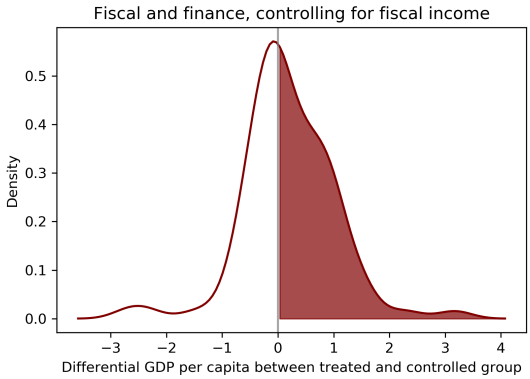


Panel C: Population and health policies

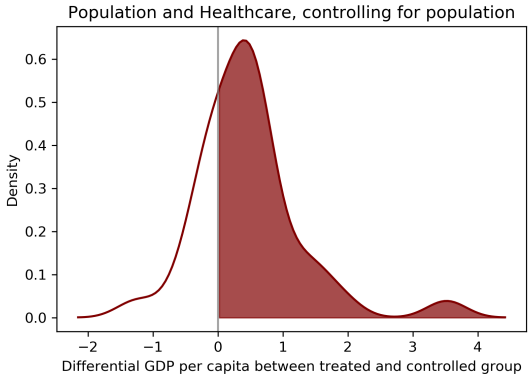
Figure A.11: This figure presents domain-specific representativeness tests. In Panel A, we focus on the subset of policies issued by the Ministry of Agriculture. We test for balance in pre-experimentation gross first-industry product. In Panel B, we focus on the subset of policies issued by the Ministry 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

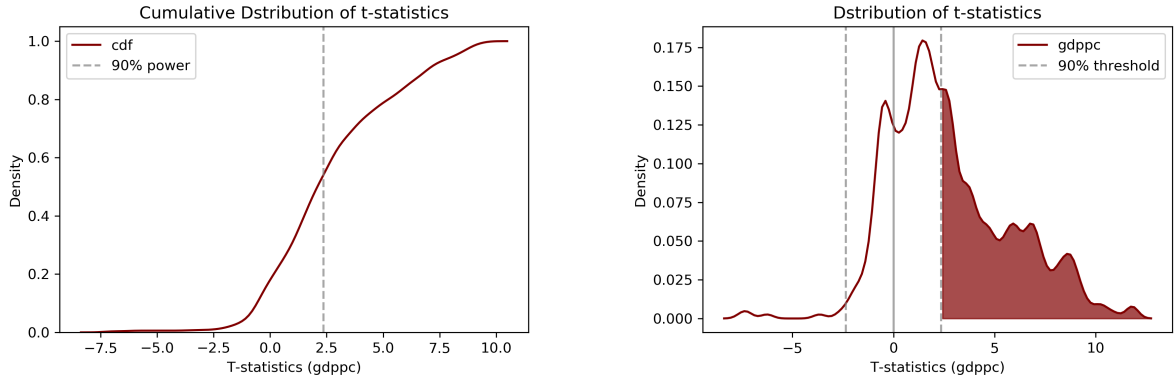


Panel B: Government finance and tax policies

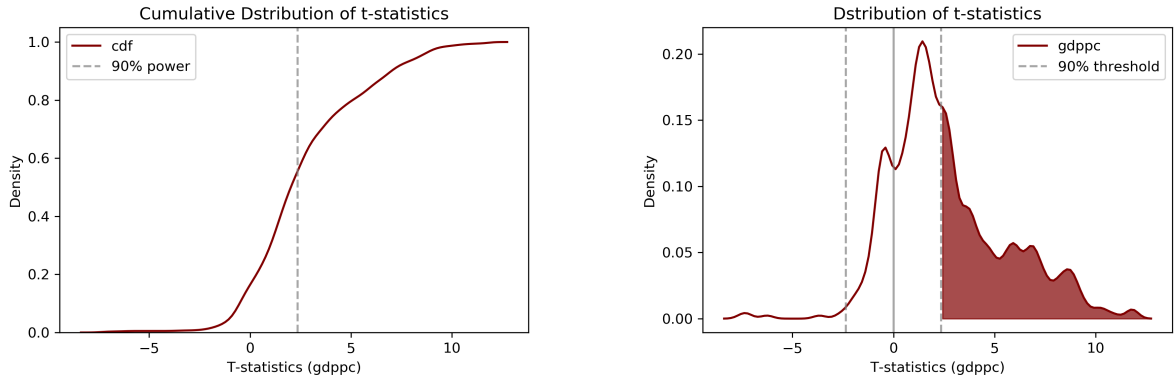


Panel C: Population and health policies

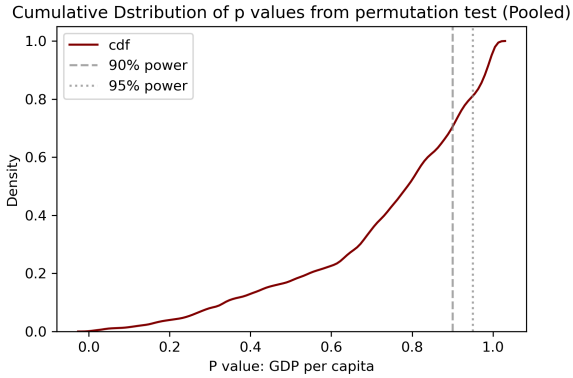
Figure A.12: This figure presents an alternative test specification where we address an assumption rationalizing positive selection by focusing on different units of analysis. We plot the distribution of regression coefficients of the treatment dummy, in a specification with controls that are believed 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.

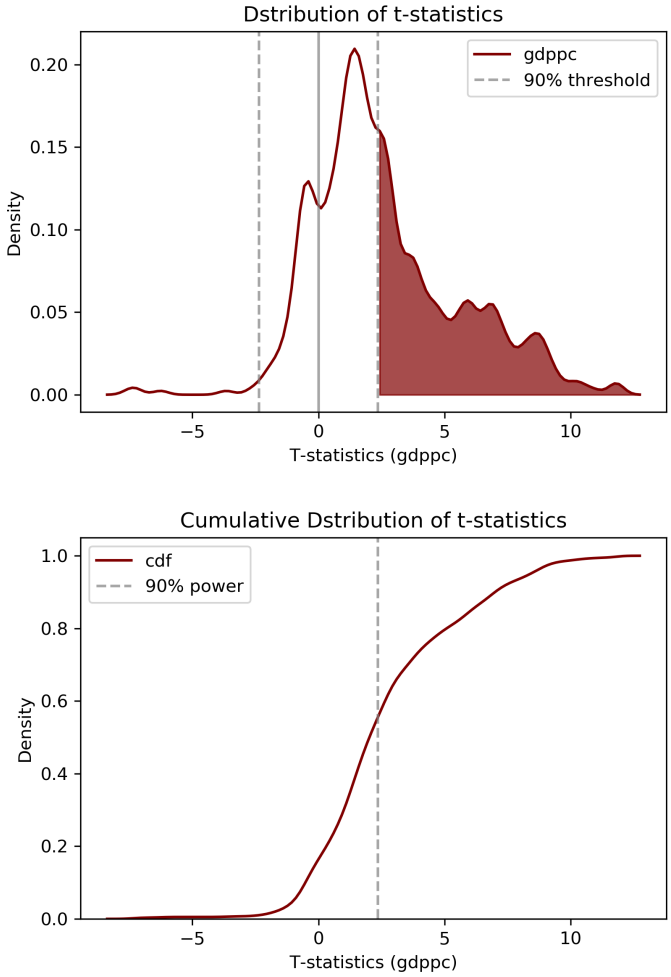


Figure A.14: This figure presents the representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. Only policies in early rounds are considered.

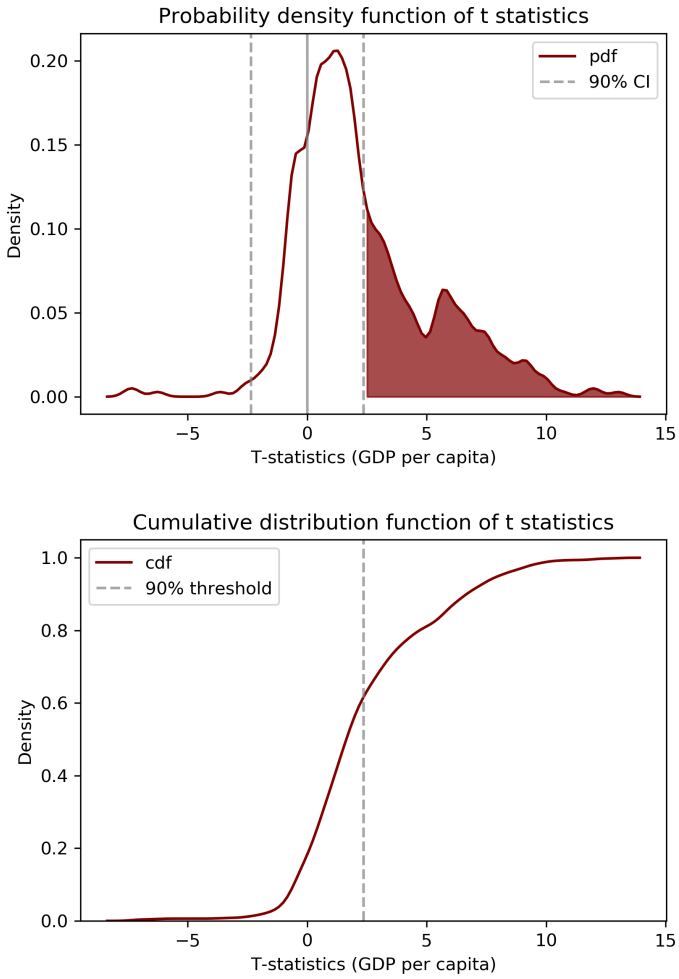


Figure A.15: This figure presents the representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. Municipalities are excluded from both treatment sample and control group

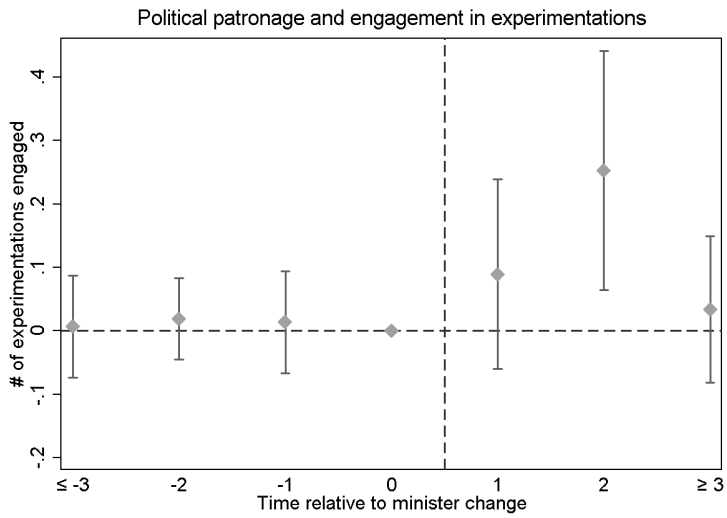


Figure A.16: This figure plots the event study estimates on a province's probability of being selected as an experimentation site after it becomes connected to a ministry due to political turnovers at the ministerial level.

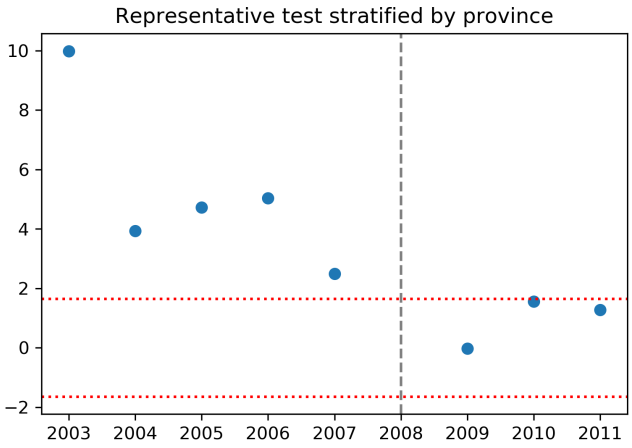


Figure A.17: Local fiscal reform — representativeness test. We conduct stratified Fisher randomization tests with student-t statistics and provincial strata. Within each province, we view counties that engage in the experimentation for the first time as units of the treatment group, the rest as control. 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.

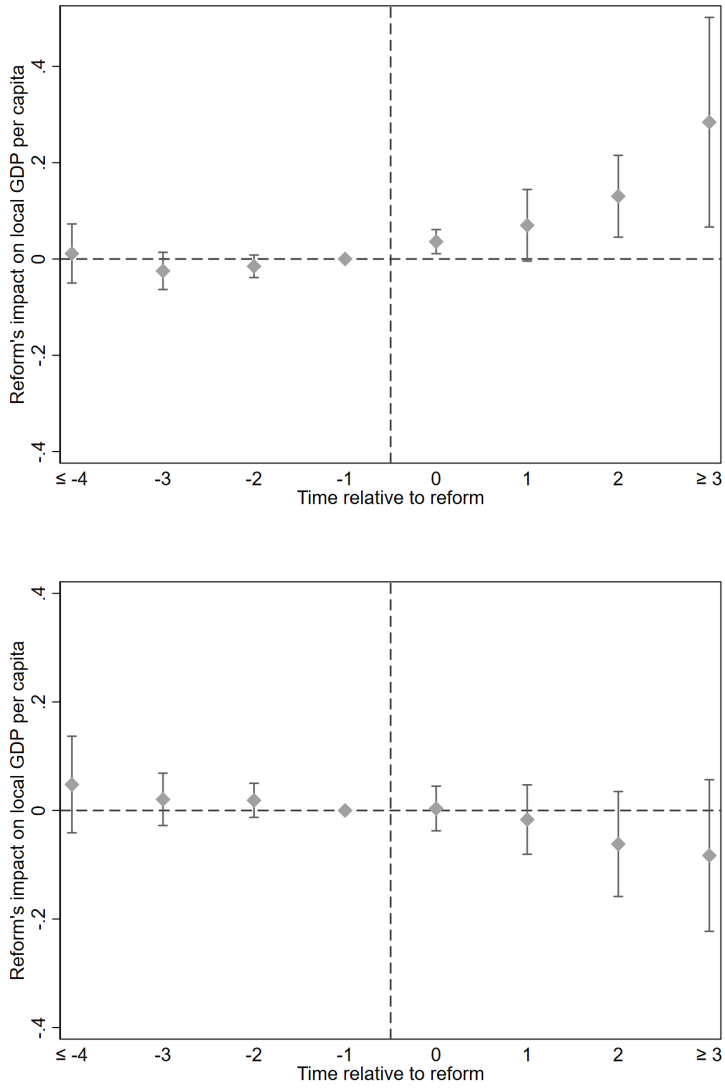


Figure A.18: Local fiscal reform - treatment effect on local GDP per capita. Upper Panel plots the counties reformed before 2007 and Lower Panel plots those reformed after 2007.

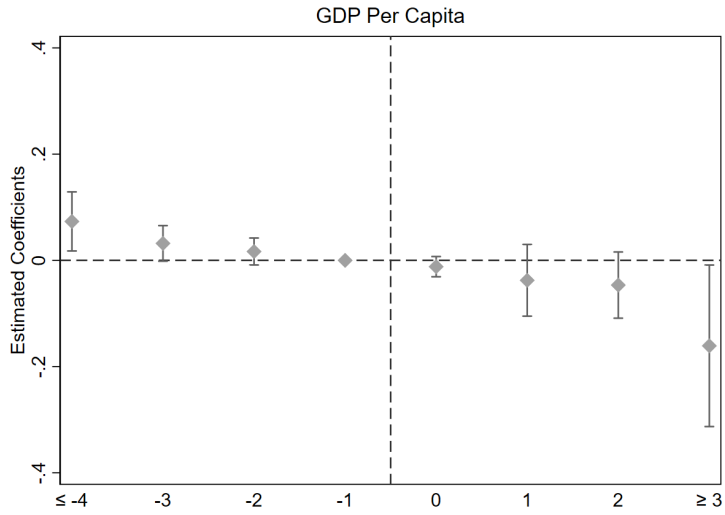


Figure A.19: Local fiscal reform - poor counties experimented before 2008. The pattern demonstrated here is similar to that in Figure A.18, Panel B.

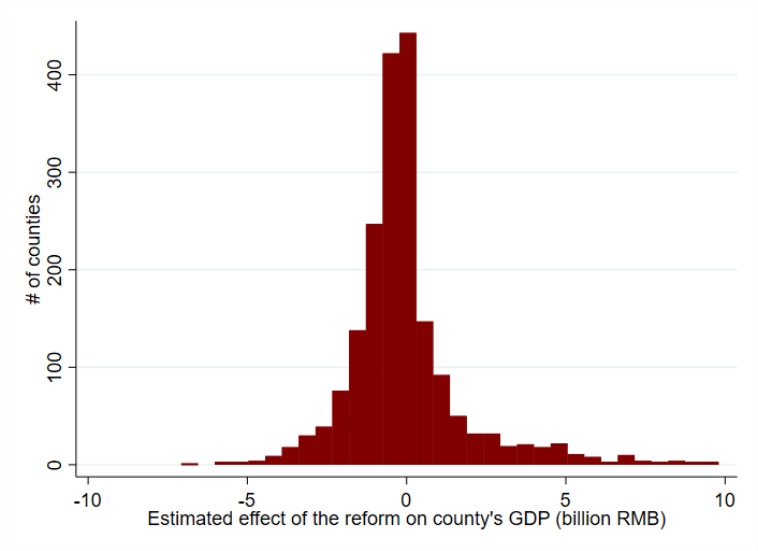


Figure A.20: Simulated treatment effects across country. We extrapolate the estimated treatment effect to all counties nationwide and obtain a distribution of reform effect on county's GDP.

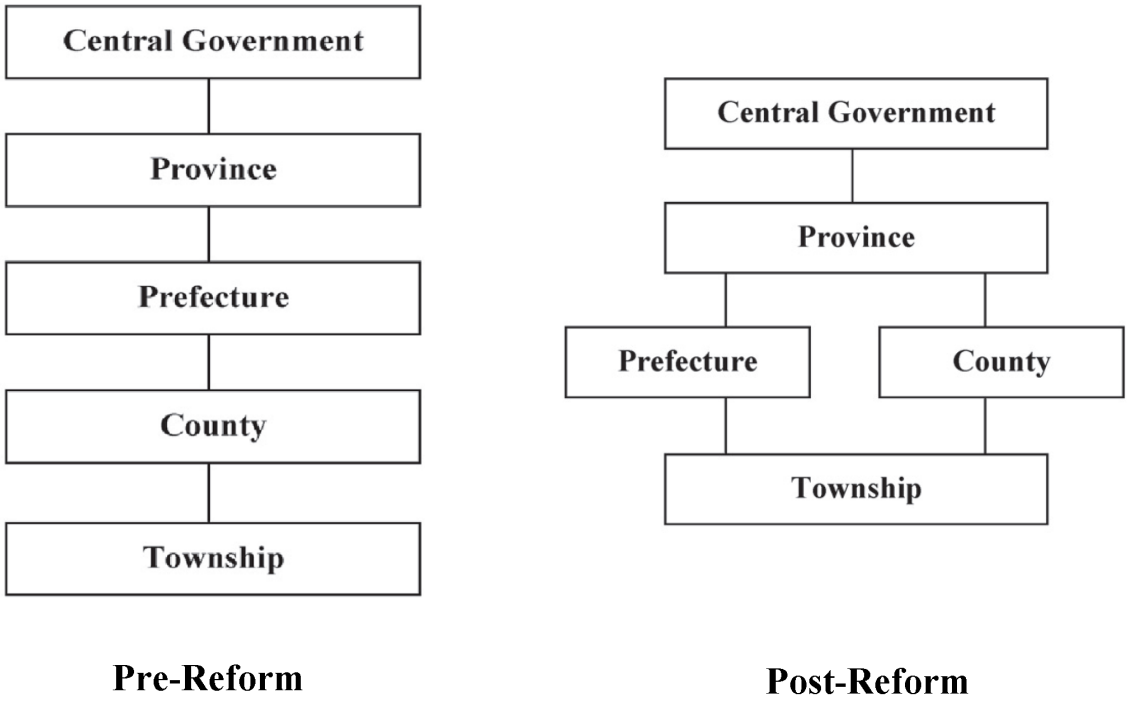


Figure A.21: Reproduced from Li, Lu, and Wang (2016). Illustration of local fiscal reform. After the reform, the provincial government could directly manage some of its counties, bypassing the prefectural cities, which grants county governments with more fiscal autonomy.

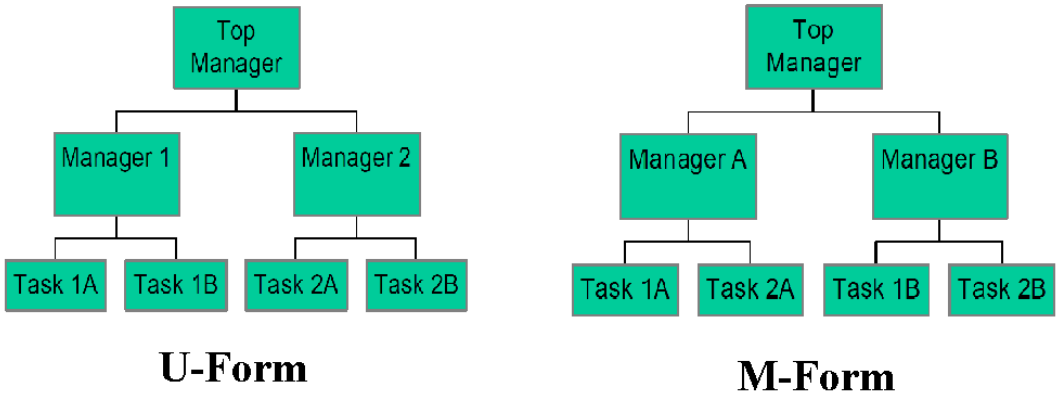


Figure A.22: Reproduced from Qian, Roland, and Xu (2006). Illustration of a shift from M form to U form. The top manager (ministers at central government branches in our case) have more administrative and personnel authority on its branches.

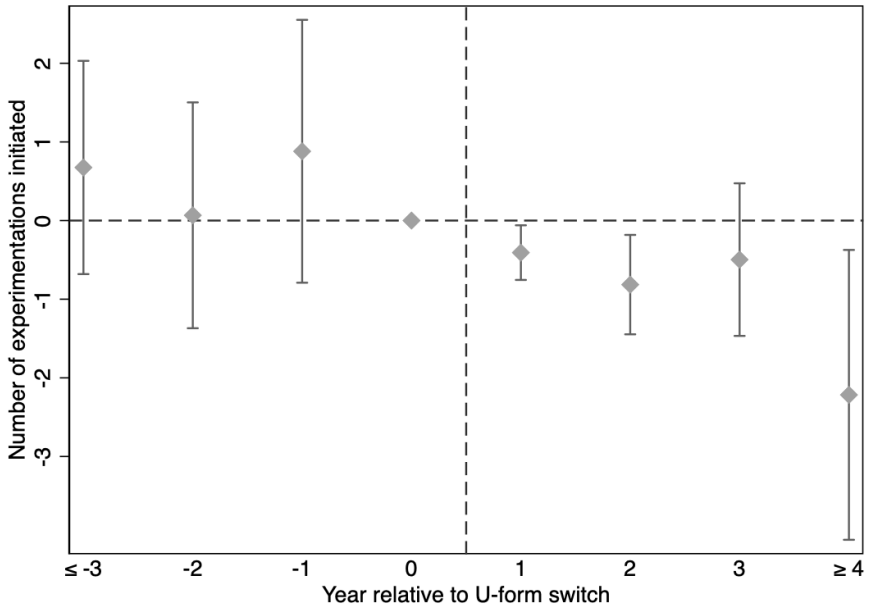


Figure A.23: Count of policy experimentations initiated after transitioning into U-form. X-axis indicates the time relative to the reform. The point estimates and confidence intervals are computed from a standard event study design controlling for ministry fixed effect and calendar year fixed effect.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

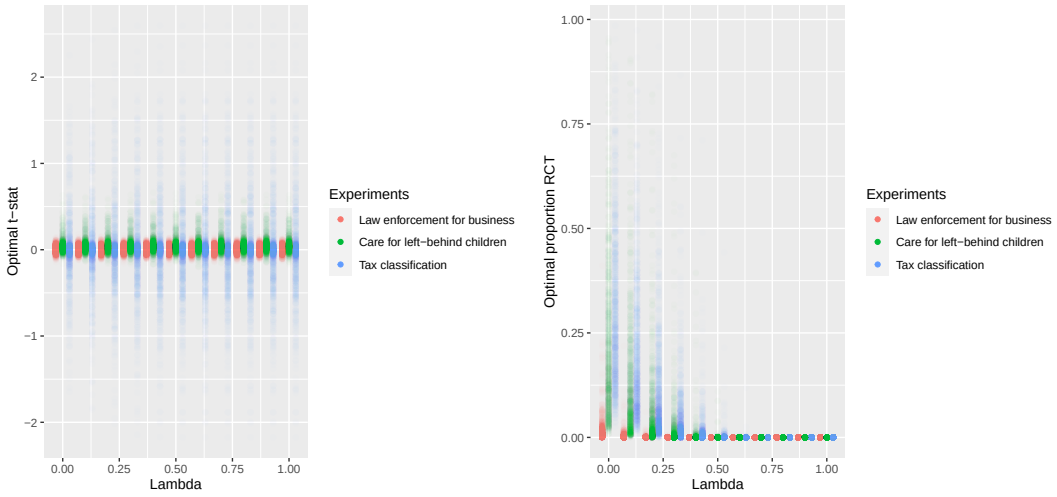


Figure A.24: This plot shows optimal t-statistics (left) and RCT vs. deterministic experimental policy (right) for simulations calibrated using three different policy experiments conducted in China following the model in Banerjee et al. 2020. Lambda ranges from 1 (full weight on decision maker's utility) 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.

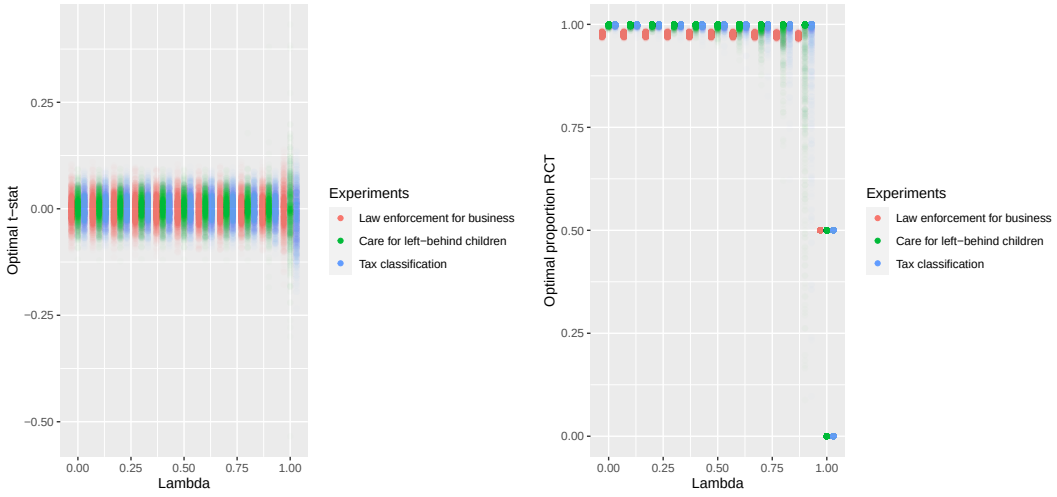


Figure A.25: This plot shows optimal t-statistics (left) and RCT vs. deterministic experimental policy (right) for simulations calibrated using three different policy experiments conducted in China following the model in Banerjee et al. 2020 with differential quality of information. Lambda ranges from 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.

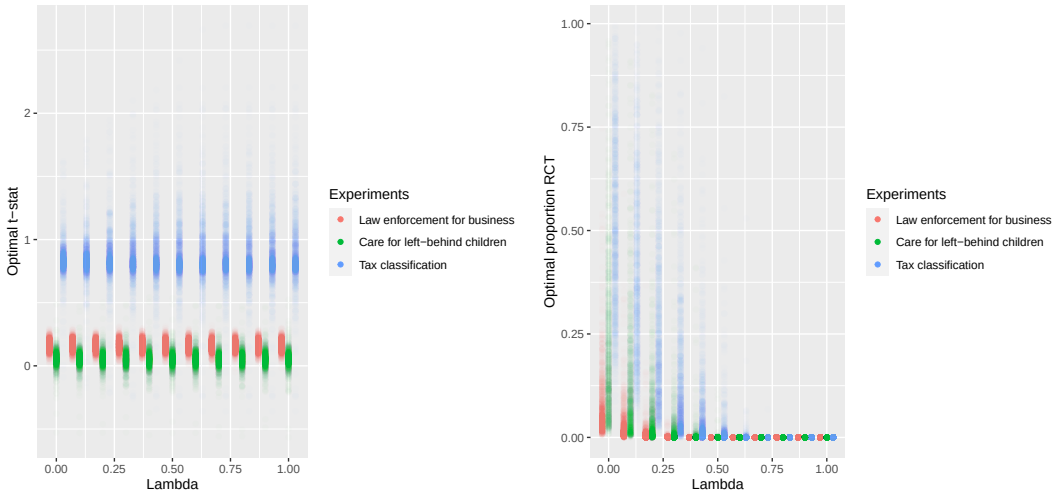


Figure A.26: This plot shows optimal t-statistics (left) and RCT vs. deterministic experimental policy (right) for simulations calibrated using three different policy experiments conducted in China following the model in Banerjee et al. 2020 with subject consent. Lambda ranges from 1 (full weight on 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.

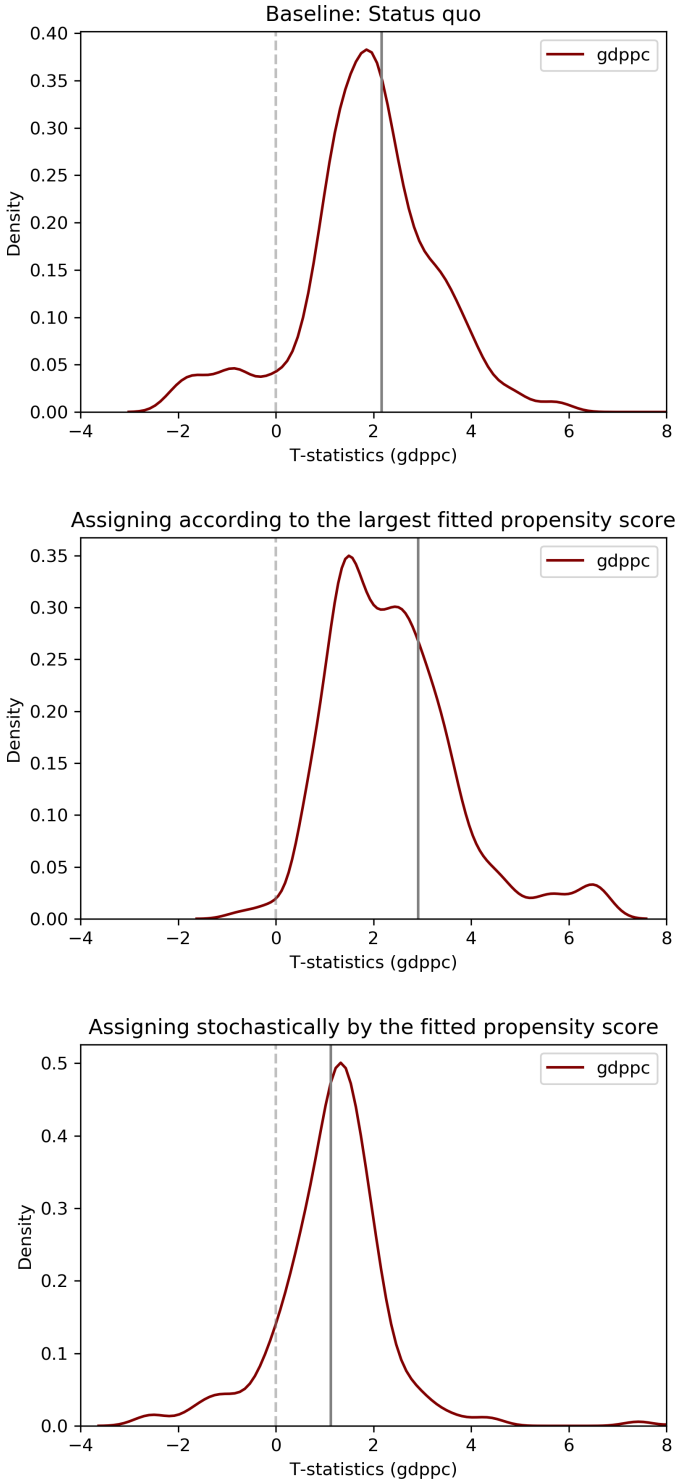


Figure A.27: Representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. To make sure we're making reasonable comparisons, we adjust the baseline Panel A so that it only includes the experimentations targeting prefectural cities, with 4 municipal 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 $\times$ 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 (1)	Full sample (2)	Certain policies (3)	Uncertain policies (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.