Policy Experimentation in China: The Political Economy of Policy Learning

Shaoda Wang
David Y. Yang*

May 21, 2024

Abstract

Many governments engage 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 describe and explain China's policy experimentation since 1980, among the largest and most systematic in recent history. We collect comprehensive data on policy experiments conducted in China over the past four decades. We document three facts. First, about 90% of the experiments exhibit positive sample selection in terms of a locality's economic development. Second, career-driven 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 central government is not fully sophisticated when interpreting experimentation outcomes. Under certain experimentation objectives, these facts imply that policy learning may be biased and national policies originating from the experimentation may be distorted. Taken together, while China's bureaucratic and institutional conditions make policy experimentation possible at an unparalleled scale, the complex political environments can also impose limitation on effective policy learning.

*. Wang: University of Chicago, NBER, and BREAD; shaoda@uchicago.edu. Yang: Harvard University, NBER, BREAD, J-PAL, and CIFAR; davidyang@fas.harvard.edu. We thank Daron Acemoglu, Isaiah Andrews, Abhijit Banerjee, Tim Besley, Mike Callen, Ting Chen, Stefano DellaVigna, Esther Duflo, Hanming Fang, Ruixue Jia, Rema Hanna, Max Kasy, Michael Kremer, John List, Mushfiq Mobarak, Ben Olken, Gerard Padro, Rohini Pande, Nancy Qian, Gautam Rao, Gerard Roland, Jesse Shapiro, Michael Song, Andrei Shleifer, Jaya Wen, Yang Xie, Daniel Xu, Li-an Zhou, and numerous conference and seminar participants for stimulating comments and suggestions. Weicheng Cai, Qingyu Chen, Andrew Kao, Jiarui Qian, Bobing Qiu, and in particular Kaicheng Luo provided outstanding research assistance.

1 Introduction

Determining which policies to implement and how to implement them is an essential task for any government (e.g., Hayek 1978; North et al. 1990). However, policy learning is challenging. The information environment that allows for assessing policy effectiveness is often complex, and factors that shape policy effectiveness are multi-faceted (including 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). Experimentation entails political and administrative procedures that allow the government to learn about novel policy instruments. Sophisticated policy experimentation has ranged from sequences of trials and errors, to pilot programs, to rigorous randomized control trials in sub-regions 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 routinely trying out policies — ranging from property tax reform, to carbon emission trading, to county fiscal empowerment reform — in a number of localities for several years before it decides whether to launch the policies in the entire nation.

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

We begin by collecting comprehensive data on policy experimentation in China between 1980 and 2020. Based on 19,812 government documents, we construct a database of 652 policy experiments initiated by 92 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, and local socioeconomic conditions. Among

these 652 policy experiments, 42.0% rolled out to become national policies after the experimentation.

We document three key facts about China's policy experimentation. First, samples of the experimentation sites are not representative. Comparing the pre-experimentation characteristics of the localities that are selected as experimentation sites and those that are not (the rest of the country), we observe that 87.7% of the experiments are conducted in sites that are positively selected in terms of local economic conditions. Experimentation sites are on average 44.2% richer in terms of local fiscal revenue than non-experimentation sites. This pattern is robust to using a number of alternative local characteristics, matching characteristics to policy domains, and implementing various testing procedures and weighting schemes.

Second, the experimental situations are not representative. In particular, we examine whether policy experimentation induces politicians' strategic efforts during the experiments. We document that local politicians participating in successful policy experiments — those leading to national policy roll-out — are substantially more likely to get promoted. In turn, local politicians exert greater effort and allocate more resources to enhance experimentation outcomes. Using a triple-differences strategy, we find that during experimentation — and not before — the ratio of local fiscal funds allocated to domains specific to the policy on trial increases by 1.3%. This is particularly the case for politicians facing stronger promotion incentives. Importantly, we find that such an increase in fiscal support is absent when the policy rolls out to the entire country, indicating that policy experiments create additional incentives and induce extra efforts that are not replicable outside of the experimentation.

Third, the central government of China is not fully sophisticated when interpreting experimentation outcomes. We find that exogenous shocks in local fiscal revenue due to unexpected land revenue windfalls during the experiments — changes to local socioeconomic conditions that are independent of policies on trial — affect decisions on whether the experimental policies roll out to the nation. Similarly, we find that routine political turnover after the experiments start — changes to local politicians' incentives that are unrelated to the nature of policies on trial — affect decisions on policies' national roll-out. Regardless of the objectives concerning policy experimentation, both of these factors during policy experiments should be discarded when evaluating experimentation outcomes.

Finally, in light of these three facts, we examine the implications for learning from experimentation and national policy outcomes. If the Chinese government is interested in learning about policies' average treatment effect (when policies are implemented to the

average locality with average local politician incentives),[1] the presence of positive sample selection and strategic efforts during experimentation could bias policy learning if the government does not fully account for these factors when making policy decisions (we provide a simple conceptual framework in Section 4, à la Al-Ubaydli, List, and Suskind (2019)). We first show that the estimator of experimentation effects that simply compares experimentation sites' outcomes before and after the experiments — thus not accounting for site selection and experimental situation — strongly predicts the trial policies' national roll-out. More sophisticated estimators, such as those using synthetic control methods, do not predict whether policies roll out nationwide. Furthermore, we find suggestive evidence that 71.1% of the policies originating from experimentation experienced shrinkage in policy effects when they rolled out to the entire country, relative to effects observed during experimentation. When a trial policy is rolled out to the entire country, localities benefit substantially more from the policy if they share similar socioeconomic conditions or if comparable local politicians share career incentives with the trial policy's 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 otherwise might prevent reform and policy adoption. On the other hand, as our results suggest, the very same institutional and bureaucratic contexts may result in deviation from representativeness in both sample selection and experimental situations (Al-Ubaydli et al. 2021; List 2022), undermining the effectiveness of policy learning from experimentation.

This paper brings an important data point to the large 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 persua-

1. In Section 8, we discuss a range of alternative objectives that may account for the patterns of policy experimentation that we observe, such as optimal experimentation design that incorporates decision makers' subjective expected utility, and experimentation structure that considers the central government's demand for political stability during and after policy experimentation.

sion across decision-makers are intertwined; and Callander and Harstad (2015) investigate how decentralized jurisdictions strategically engage in policy experimentation, and how a central government can encourage policy convergence. Closest to our context, 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 with the first empirical analyses of the comprehensive set of policy experiments that have been conducted in China over the past four decades. While China's policy experimentation is one of the largest systematic policy learning institutions in history, surprisingly little is known about its characteristics and how it affects China's policy landscape. We highlight that specific institutional contexts shape the structure of experiments and affect their outcomes.[2]

Our work also adds to the growing literature on policy learning and policy scale-up, especially the recent studies highlighting the structural factors that may limit how policy trials can inform broader outcomes after pilot programs are scaled up (e.g., Al-Ubaydli et al. 2017; Davis et al. 2017; Al-Ubaydli, List, and Suskind 2019; Al-Ubaydli et al. 2021; List 2022). Consistent with the theoretical framework proposed by (Al-Ubaydli, List, and Suskind 2019), we find that both non-representative experimental samples and non-representative experimental situations could be key reasons for the lack of scalability. Moreover, we find that policymakers do not fully account for characteristics of the experimental sample and situation, and are thus unable to predict whether experimental findings will end up being "scalable." The patterns we document include positive experimentation site selection in general, and, in particular, diminishing policy effects as the policy is expanded beyond the experimentation sites, which have better socioeconomic conditions and extra political incentives. These patterns echo similar findings by Allcott (2015) on the sample selection bias in the Opower energy conservation programs in the U.S., as well as findings by DellaVigna and Linos (2020) thon trials conducted by the Nudge Units in the U.K. had smaller effects when scaled up, due to changes in the intervention, institutional contexts, and implementation details. Our findings also are 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.[3]

2. Related literature has attributed China's success with economic decentralization to its powerful political centralization (Blanchard and Shleifer 2001; Xu 2011), which fosters competition for promotion among local politicians on dimensions aligned with the central government's policy goals (e.g., Li and Zhou 2005; Jia, Kudamatsu, and Seim 2015; Bai, Hsieh, and Song 2020). Our results complement this literature by highlighting a classic pitfall of political centralization due to incomplete contract (Kornai 1959; He, Wang, and Zhang 2020).

3. Intriguingly, these patterns stand in contrast with the limited positive selection among the US states that are leaders in policy innovations (DellaVigna and Kim 2022); they also contrast with the limited site

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 experiments, 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 when processing information. They may not internalize information acquisition costs due to political hierarchy (Rogger and Somani 2018). They may fail to take into account the context of the study (Hjort et al. 2021) or 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 outlines a simple framework on policy learning and factors that may affect outcomes of policy learning, which organizes the subsequent empirical analyses. The following three sections present the three key facts on policy experimentation: sample selection of experimentation sites (Section 5), strategic efforts by local politicians during the experiments (Section 6), and non-sophisticated interpretation of experimentation outcomes (Section 7). These can be interpreted without taking a stance on the central government's objectives for experimentation. Section 8 discusses the implications for learning from experimentation and national policy outcomes, under specific assumptions about the government's objectives. Finally, Section 9 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,

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). 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).

powerful 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 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 (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 models, the experiments during the reform and opening-up era are distinguished by their open-endedness in generating novel instruments and solutions. The "institutional entrepreneurship" unleashed 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 can 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 usually announces and introduces 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 also 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).[4]

Once a policy experiment is determined to be successful, certain experimentation points are set as demonstration zones (*Shifanqu*). 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

4. 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.

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 revoked.

In this paper, we focus primarily on policy experiments through experimentation points, including those clustered in experimental 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.

3 Data and characteristics of policy experimentation

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 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, of which 4,399 were issued by the central government and 15,413 by local governments. Central government docu-

ments 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 implementation and administrative arrangements.

We identify 652 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.[5]

Among the 652 experiments, 613 involve policies explicitly intended for potential national roll-out, and 39 are policies with specific regional targets.[6] For the baseline analysis in Section 5, where we examine experimentation site selection, we exclude policies with explicit regional targets; however, the results are robust if we include all policy experiments in the analyses and adjust the sampling frame according to the specific experiments' regional scope. We exclude 109 policy experiments that are still ongoing when we examine whether the policies on trial have been rolled out to the whole country (throughout Sections 6- 8)

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

We conduct various cross-checks to ensure the comprehensiveness of the government documents that we collect. For the ministries that publish documents on their own websites, we independently collect documents from the ministerial websites. We find that *PKULaw.com* has extensive and comprehensive coverage (see Appendix Table A.2). When

5. For example, experiments on corn seed insurance, rice production insurance, professional farmer training, and agricultural technology promotion and consulting are combined into an overarching experiment on improving agricultural technology and management. Our results do not qualitatively change if we undo the grouping and treat the experiments as independent trials.

6. 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.

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 additional critical information.

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.[7] 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 fails 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* mean that policy uncertainty is irrelevant in this context; on average, 58.0% 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 1 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 1980s and 1990s, averaging less than 10 new experiments per year across all ministries and commissions. The number of experiments began to increase sharply toward the end of the 1990s, reaching a peak of 47 experiments initiated in 2010 alone, and has gradually declined since then.

While many factors could contribute to these patterns, part of the decline 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 local governments to upper-level ministerial units, they move away from flat,

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

multi-divisional structures (M-form), which may provide flexibility and ease in coordinating policy experiments, to more centralized, unitary structures (U-form), which 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 We extract the experimentation sites of each policy experiment.[8] Figure 2, 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 experiment. In addition, 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 42.6% of the experiments allow (at least partially) for voluntary participation of the local governments (see Appendix Figure A.2).

National policy 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, 42.0% of the policy experiments eventually became national policies, while 58.0% failed (see Figure 1, 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 policies' national roll-out are not sensitive to the particular definition: we alternatively define an experimental policy as being rolled-out nationally if the experiment ends up covering at least two-thirds of the provinces, and we find similar patterns (see Appendix Figure A.4).

Policy domains and involved ministries We identify all the central government ministries and commissions involved in a policy experiment, and measure each ministry or

8. Many policy experiments have more than one wave of roll-out, and we identify 1,374 distinct rounds of roll-out across the 652 experiments. In this paper, we pool all rounds together. 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. We leave it to future work to systematically study the dynamic experimentation implementation.

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 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.5 plots the count of policy experiments by policy domain over time.

Uncertainty and complexity We construct a number of measures for the *ex ante* uncertainty of each policy experiment. We a consider policy on trial to be more certain based on several criteria: *(i)* if the central government has laid out a detailed national roll-out timeline before the experiment starts;[9] *(ii)* if experimentation details were already drafted out by the central government at the beginning of the experiment; or *(iii)* if the policy was mentioned in the Five-Year Plans, signaling greater political will to make the policy national. We also construct a measure for academic consensus of the policy on trial, where we match each policy to academic papers published prior to the beginning of the experiment; we calculate the average textual similarity across these papers (using TF-IDF).

We also construct a number of measures for the complexity of each policy experiment. We consider a policy on trial to be more complex based on the following criteria: *(i)* if multiple ministries and commissions are involved in the experiment;[10] *(ii)* if the government documents describing the experiments are long and/or contain multiple documents; *(iii)* if the experiment duration is long; or *(iv)* if there are a large number of relevant local government documents that complement the central government document.

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 categorize whether the policy experiment is aiming at relatively short-term outcomes based on the time frame described in the experimentation documents. We identify whether the central government provided additional fiscal support for the experimentation sites, whether the policies on trial would in principle benefit from extra fiscal support, and how the local government would allocate fiscal resources to policy domains related

9. 30.8% of the experiments feature such timelines (which we label as experiments on policies with high certainty), and 61.9% of them eventually become national policies. In contrast, among the 69.2% of experiments that do not feature such a timeline (which we label as experiments on policies with high uncertainty), only 35.6% were eventually rolled out to the entire country (see Appendix Figure A.6).

10. 23.8% of the policy experiments involve more than two ministries and commissions; we label these as complex experiments (see Appendix Figure A.7)

to the experiment. We also measure policy differentiation across time and space, by constructing matrices of pairwise textual similarities for all the local policy documents that belong to the same policy experiment.[11]

3.3 Four examples of policy experimentation

We map four distinct policy experiments to illustrate the range of policy experimentation that took place in recent decades (see Appendix A.2 for additional details of those examples). In addition, in Appendix Table A.1, we present examples of policy experiments across a variety of policy domains.

Figure 2, Panel B.1 depicts the experimentation on carbon emission trading, 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 that aims to reduce 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 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 toward inland, less developed regions. The experimentation ended in 2015 and the fiscal empowerment reform did not roll out to the country.

11. Text similarity is calculated by a pre-trained model from *paddlepaddle*, and we provide more details in Section G

4 Conceptual framework

To guide our empirical analyses, we now describe a simple conceptual framework — following Al-Ubaydli, List, and Suskind (2019) — that highlights the key factors that may influence policy learning during policy experimentation.

We denote observed experimentation outcomes as $\hat{Y}(p, \mathbb{I}_t, \mathbb{E}_t)$, where Y is the policy outcome of interest,[12] p corresponds to specific policy of interest, $\mathbb{I}_t$ the pre-experimentation socioeconomic characteristics of localities where the policy experiment takes place, and $\mathbb{E}_t$ represents the local politicians' incentives and efforts during policy experimentation.

We can decompose the observed experimentation outcomes as follows:

$$\hat{Y}(p, \mathbb{I}_t, \mathbb{E}_t) = \underbrace{\bar{Y}(p, \bar{i_t}, \bar{e_t})}_{\text{ATE}} + \underbrace{F_{i,p}(\mathbb{I}_t - \bar{i_t})}_{\text{Site selection}} + \underbrace{F_{e,p}(\mathbb{E}_t - \bar{e_t})}_{\text{Experimental situation}} + G_{i,e}(\mathbb{I}_t, \mathbb{E}_t), \tag{1}$$

where $\bar{Y}(p, \bar{i_t}, \bar{e_t})$ indicates the average effect of policy p when it is implemented in localities with the average socioeconomic characteristics ($\bar{i_t}$) and the average local politicians' incentives and efforts ($\bar{e_t}$). The average treatment effect may be a parameter of interest to the policymaker because it indicates the expected outcome of the policy on trial when it's rolled out to the whole country.

While the observed experimentation outcome $\hat{Y}$ is a function of $\bar{Y}$, the two can differ due to a number of factors. First, to the extent that policy effects are often heterogeneous across localities, $\hat{Y}$ and $\bar{Y}$ can diverge if the selection of experimentation sites is not representative of the average locality. For example, policies that achieve favorable outcomes in rich regions during experimentation do not necessarily generate comparable outcomes when they subsequently roll out to poor regions. $F_{i,p}(\mathbb{I}_t - \bar{i_t})$ captures the heterogeneous policy effects with respect to localities' *ex ante* socioeconomic characteristics. Section 5 documents non-representative selection of experimentation sites, that is, $\mathbb{I}_t - \bar{i_t} \neq 0$.

Second, to the extent that efforts of the key actors (i.e., local politicians) can play significant roles in shaping policy outcomes, $\hat{Y}$ and $\bar{Y}$ can diverge if the experimental situation that induces local politicians' efforts is not representative. For example, policy experiments may generate excessive efforts among local politicians because they consider favorable experimental outcomes a salient signal to the central government and a significant contributor to their career advancement. $F_{i,p}(\mathbb{E}_t - \bar{e_t})$ captures the heterogeneous policy effects with respect to local governments' effort during implementation. Section 6 aims to document the presence of non-representative experimental situations, that is, $\mathbb{E}_t - \bar{e_t} \neq 0$.

12. In our analysis, we use local GDP/fiscal growth, which captures the primary policy incentives for the Chinese government (Li and Zhou 2005).

Furthermore, we note that $\hat{Y}$ and $\bar{Y}$ can diverge if experimentation sites' socioeconomic characteristics and local politicians' incentives are associated with outcomes of interest. This could occur either due to factors unrelated to the policy on trial, or through the policy on trial but in an indirect or unintended manner. For example, good rainfall may boost local economic growth during the year of experimentation, but this has nothing to do with the policy being tried. We denote this as $G_{i,e}(\mathbb{I}_t, \mathbb{E}_t)$; this term does *not* depend on the policy p.[13]

Given the observed experimentation outcome, we denote the central government's decision rule (D) on whether to roll out a policy nationwide as follows:

$$D\left(\hat{Y}(p, \mathbb{I}_t, \mathbb{E}_t) - \delta_1(F_{i,p}(\mathbb{I}_t - \bar{i}_t) + F_{e,p}(\mathbb{E}_t - \bar{e}_t)) - \underbrace{\delta_2}_{\text{Naivete}} G_{i,e}(\mathbb{I}_t, \mathbb{E}_t) + \delta_3 \mathbb{Z}_{pit}\right). \quad (2)$$

As the central government evaluates the policy experimentation outcomes ($\hat{Y}(p, \mathbb{I}_t, \mathbb{E}_t)$) and decides whether to roll out the policy to the entire nation, the decision depends on the average policy effects that one may infer from the experimentation outcomes ($\bar{Y}(p, \bar{i}_t, \bar{e}_t)$). Importantly, we also allow for the possibility that the decision rule incorporates the non-representative sample selection and non-representative experimentation situation components of the experimentation outcomes. If the central government wishes to learn about policies that maximize the average policy effects, then $\delta_1 = 1$, because the government should fully account for the role of site selection and non-representative situation in affecting experimentation outcomes. When $\delta_1 \neq 1$, it may reflect the government's lack of sophistication if its objective is to learn about policies that maximize the average policy effects; it may also capture the deviation of experimentation objective away from maximizing average policy effects (for example, an objective function that gives more weight to the economic performance of certain localities in the country).

The term δ_2 captures the possibility that the presence of non-representative sample selection and non-representative experimentation situations could affect the central government's policy decision, even if they are independent of the policy on trial ($G_{i,e}(\mathbb{I}_t, \mathbb{E}_t)$). This is different from the previous term: if $\delta_2 \neq 1$, it explicitly indicates the central government's lack of (full) sophistication when evaluating experimentation outcomes, which cannot be explained by alternative experimentation objectives. Section 7 aims to document that the central government is indeed not fully sophisticated when interpreting ex-

13. $\hat{Y}$ and $\bar{Y}$ may diverge due to a range of other factors, such as general equilibrium effects, which could either amplify or shrink the policy effects when the policy is implemented in a small number of localities versus the entire nation. This is beyond the scope of our empirical investigation; hence, we do not explicitly include these factors in the conceptual framework.

perimentation outcomes, and fails to discount experimentation sites' characteristics and politicians' strategic incentives, which are correlated with the underlying outcomes but independent of the policy on trial.

Finally, $\mathbb{Z}_{pit}$ denotes outcomes other than economic performance that occurred during experimentation (e.g., local political stability). The term δ_3 captures aspects of policy learning from experimentation beyond policy effects on economic growth (e.g., to minimize the prospect of local unrest). While not the main focus of our paper, we discuss these other considerations in Section 8.

5 Is the selection of experimentation sites representative?

In this section, we ask whether the selection of experimentation sites is representative of China's localities. In the language of the framework presented in Section 4, we test whether $\mathbb{I}_t - \bar{i}_t = 0$.

5.1 Procedure to test for representativeness

For each policy experiment, we compare pre-experimentation characteristics between localities that participate in the experiment and those that do not. As the baseline, we examine the local fiscal expenditure in the year before the experiment begins, and we conduct t-tests against the null hypothesis that the pre-experimentation levels of local fiscal expenditure are indistinguishable among the experimentation sites and non-experimentation sites. This amounts to 652 separate t-tests, one for each policy experiment.[14] In Section 5.2, we describe a range of alternative tests and definitions of representativeness.

We use the corresponding t-statistics as summary statistics to quantify the deviation from representativeness for each policy experiment. The *student's-t* statistic for policy experiment i is:

$$t_i = \frac{\hat{Y}_i(1) - \hat{Y}_i(0)}{\sqrt{\frac{\hat{S}_i^2(1)}{n_{i,1}} + \frac{\hat{S}_i^2(0)}{n_{i,0}}}}, \tag{3}$$

following the t-distribution with degrees of freedom ν_i.[15]

14. 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.

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

The specific context of China's policy experimentation poses two complications in conducting these representativeness tests. First, policy experiments can be implemented at the provincial, prefectural, or county level. We conduct the representativeness tests at the appropriate administrative level for each policy experiment. The county and prefectural level experiments 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 jurisdiction to implement the experiment. Thus, 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.[16]

Second, approximately one-fourth of the experiments involve only one experimentation site. We cannot conduct standard statistical tests for these single-site experiments. Instead, we pool each single-site experiment with four other randomly selected single-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 yields a corresponding t-statistic for each of the one-site experiments. In addition, we conduct a range of alternative tests concerning these one-site experiments, such as pooling experiments that take place in consecutive periods, and drawing bootstrap samples with replacement.

5.2 Most experimentation sites are positively selected

In Figure 3, Panel A, we plot the distribution of the baseline t-statistics comparing pre-experimentation local fiscal revenue between the experimentation and non-experimentation sites. We mark the thresholds of t-statistics where one would reject the null hypothesis of representative site selection at the 95% confidence interval.[17] Table 1, Panel A, reports the corresponding test statistics (adjusting for the degrees of freedom for each test) and the share of policy experiments for which we can reject the null hypotheses at the 5% significance level (in the last two columns).

We find that the average of the t-statistics comparing experimentation and non-experimentation

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

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

17. As discussed above, each of the 652 t-tests has its specific degrees of freedom. We depict visually the average width of the 95% confidence interval (3.33).

sites is 5.17, across all experiments that are intended for national roll-out. For 87.7% of the experiments, experimentation sites are on average richer than localities that do not participate in the corresponding experiments. Applying statistical tests that are fairly conservative, we are able to reject the null hypothesis of representative site selection among 57.0% of the experiments at the 5% level.[18]

The positive selection of experimentation sites is a robust pattern. We assess the robustness across various test samples and socioeconomic characteristics used for the test. Regarding the test samples, in addition to the baseline specifications where we focus on all experiments that intend for national roll-out, we: *(i)* include experiments on policies that target specific regions and adjust the non-experimentation sites according to the regional scope; *(ii)* focus on just the initial round of experimentation states participating in a given experiment if there are multiple rounds; *(iii)* exclude the selection of centrally-administered municipalities such as Beijing and Shanghai, where local economic development and the central government's priorities for policy implementation may coincide; and *(iv)* construct a one-site experiment sample by pooling other one-site experiments taking place around the same year. Regarding socioeconomic conditions that are compared between localities participating in the experiments and those that are not, we focus on a number of alternative dimensions measured before the experiments start: *(i)* local GDP; *(ii)* local GDP per capita; *(iii)* local GDP growth rate during five years prior to the experiment; *(iv)* local population; and *(v)* local fiscal expenditure. In Figure 3, Panel B, we plot the average t-statistics comparing experimentation and non-experimentation sites, using all combinations of the variants of sample and testing characteristics. We continue to observe positive t-statistics throughout all tests.

Furthermore, we take into account the different policy domains across experiments when conducting tests for experimentation site selection. First, we break down experiments and tests for representative site selection by each policy domain (Table 1, Panel B, reports the summary statistics; Appendix Figure A.9, Panels A to N, plot the distribution of the t-statistics). We find similar (if not starker) patterns of positive experimentation site selection. Second, we match experiments in different policy domains with the domain-specific pre-experimentation characteristics and replicate the test for representativeness (Average $t_{argriculture} = 2.48$, $t_{fiscal} = 5.26$, $t_{population} = 2.61$, see Appendix Figure A.10, Panels A to C). We continue to find strong patterns of positive selection. For example, agricultural policy experiments take place in localities with substantially higher

18. The average difference between experimentation sites and non-experimentation sites is 961 million Yuan in terms of local GDP (26.0% of the average non-experimentation site GDP), 731 Yuan (10.1%)in terms of GDP per capita, 42.5 million Yuan (31.8%) in terms of local fiscal revenue, and 4.75 million Yuan (19.1%) in terms of domain-specific local fiscal expenditure.

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. Third, 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). Fourth, we classify policies into pro-poor (involving rural regions or targeting a poor population; constituting 42% of the sample) and pro-rich (targeting general economic development) and we separately examine the positive selection within each category (see Appendix Figure A.10, Panel E). Pro-rich policies indeed exhibit more positive site selection than pro-poor policies; however, even the the subset of pro-poor policies has experimentation sites that are significantly positively selected. Fifth, we consider the possibility that administrative localities may not be the natural unit of analysis when policies in certain domains are evaluated nationally. Specifically, we replicate the baseline test for representative site selection, while weighing localities by their rural population size in the case of agricultural policy experiments (see Appendix Figure A.10, Panel F), by the total number of firms in the locality in the case of experiments with government finance and tax policies (see Panel G), and by the total population size in the case of experiments with population and health policies (see Panel H). These weighted t-tests continue to exhibit a substantial mass of t-statistics above zero.

Finally, the pattern of positive selection is robust to a wider family of statistical tests, such as using permutation tests (see Appendix Figure A.11).

5.3 Potential reasons for observed positive selection

Having documented that the selection of experimentation sites is not representative, we now provide several explanations that may explain such positive selection.

There may be stronger positive site selection for policies about which the central government is fairly certain; in those cases, learning might not be the most important objective for the policy experiments. As described in Section 3, we measure *ex-ante* uncertainty for each policy experiment using proxies such as whether the central government already has laid out detailed national policy roll-out plans at the beginning of the experiment, and the level of consensus exhibited by academic publications before the experiment. Appendix Table A.3, Panel A, presents the correlation between the baseline t-statistics and each proxy for *ex-ante* uncertainty of the corresponding experiment (and an index summarizing all proxies). We find that, contrary to the hypothesis, experiments that show

signs of more certainty of national roll-out are associated with a *weaker* degree of positive site selection.

Another possible explanation for positive site selection is that, for policies that are complicated to implement, richer localities with stronger local governance capacity may provide more precise signals on policy effectiveness. As described in Section 3, we measure policy complexity using proxies such as whether multiple ministries are involved in the policy and the length of the description of the experiment. Appendix Table A.3, Panel B, presents the correlation between the baseline t-statistics and each proxy for complexity of the corresponding experiment (and an index summarizing all proxies). We find that, consistent with the hypothesis, more complex policies tend to be associated with *more* positive selection in the experimentation sites.

Yet another explanation could be misaligned interests between the central and local governments. 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 criterion 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 institutional innovation at the national level. [...] Hence, the experimentation sites should be fairly representative.

We indeed observe that provincial level policy experiments, whose experimentation sites are directly selected by the central government, are much less positively selected on average (see Table 1, Panel C). In contrast, policy experiments at the prefectural and county levels, whose experimentation sites are often selected by provincial governments conditional on their being selected as experimentation provinces, are substantially more positively selected. This suggests that, while the central government may be concerned about policy learning, the local governments may not fully internalize such objectives. We examine local officials' career incentives more explicitly in Section 6.

There are many potential reasons that could cause positive selection in experimentation sites, and we do *not* intend to pin down the exact mechanisms behind the observed deviation from representativeness. Regardless of the source, if the central government does not take positive selection into full account when evaluating experimentation out-

comes, then it could affect policy learning.[19] We examine the implications on policy learning and national policy outcomes in Section 8.

6 Do experiments induce strategic efforts?

In this section, we ask whether the experimental situations are representative, in particular, whether the policy experimentation induces strategic efforts among participating local politicians. In the language of the framework presented in Section 4, we test whether $\mathbb{E}_t - \bar{e}_t = 0$.

We begin by documenting the link between politicians' career incentives and their participation in policy experimentation. In particular, we focus on promotion within the political hierarchy, which is a central objective that motivates local politicians (Li and Zhou 2005, Jia, Kudamatsu, and Seim 2015, Jiang 2018). For each local politician (party secretary), we predict the likelihood of promotion based on the number of policy experiments in which she has participated during her tenure as a local leader; we control for locality fixed effects and position term fixed effects.

Appendix Table A.5, columns 1-4, presents the results. We find that, while participation in experiments *per se* is not associated with political promotion, being part of successful experiments — those eventually rolled out to the entire country as national policies — during one's tenure is associated with a substantial increase in local politicians' promotions. Having participated in one successful experiment corresponds to a 23.5% increase in the probability of promotion. As columns 5-6 show, such association is stronger if the experiments are small-scale (those with fewer than 10 experimentation sites), consistent with the hypothesis that the reward from a successful policy trial is shared among fewer competing politicians.

These correlational patterns suggest that policy experiments — due to their high visibility and high political reward (only when they end up leading to national policies) — may induce local politicians to exert greater efforts to achieve successful outcomes during an experiment, in order to increase the chance that the policy on trial will roll out to the entire country.

19. We observe a modest decrease in the positive selection of experimentation sites over the years, suggesting that the central government may have learned and corrected for the positive selection, albeit very mildly. Appendix Figure A.8 plots the overall share of positively selected experiments over the four decades since 1980, and Appendix Table A.4 presents regression results on the time trend in positive selection, for all experiments and separately by ministry.

6.1 Allocation of fiscal resources during experimentation

Local fiscal expenditure is an important input in policy outcomes. To the extent that local politicians may be rewarded for successful policy experiments, do local governments participating in such experiments significantly increase fiscal expenditure, which may improve the experimentation outcomes?

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.[20] We then use a triple-differences strategy to examine whether the start of policy experimentation in a specific domain causes increases in fiscal expenditure in the corresponding domain, relative to the general trend of domain-specific expenditure in a given county and in a given year. Specifically, we estimate the following 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; and Exp_{ikt} is the number of experiments in fiscal domain k that county i engaged in during year t.[21] We include full sets of county-by-year fixed effects (λ_{it}), domain-by-year fixed effects (δ_{kt}), and county-by-domain fixed effects (δ_{kt}), which allow us to isolate changes in local politicians' behaviors due to policy experiments in a specific domain that started in a specific year in certain localities. The standard errors are clustered at the county level.

The results are presented in Table 2, Panel A, columns 1-3. We observe a significant increase in domain-specific fiscal expenditure: an additional experiment increases local expenditure in the corresponding domain by about 1.3% in terms of share of total fiscal expenditure.[22]

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). Politicians' career incentives are measured as a combination of their starting age of tenure and bureaucratic rank, following Wang, Zhang, and Zhou (2020).[23]

20. They are: general administrative cost, infrastructure, economic production, agriculture / forestry / fishing, science / education / culture / health, and others.

21. For 96.8% of the observations, the number of experiments is either 0 or 1.

22. Local fiscal expenditure data (along with fiscal revenue) is among the least manipulable information due to its double book-entry nature (Jia, Guo, and Zhang 2014). Thus, the increased local fiscal expenditure is unlikely to reflect data manipulation or exaggerated reports of local socioeconomic performance.

23. Specifically, we collect detailed biographical information on the universe of Chinese ministers and provincial/prefectural leaders during our four-decade sample period, and estimate each prefectural city leader's ex ante likelihood of promotion in each year, as a flexible function of his age when starting the

This heterogeneity is consistent with the hypothesis that politically incentivized local leaders are particularly keen on making sure the policy experiments succeed in their jurisdictions.[24] Moreover, we observe that local politicians' fiscal reallocation toward the domains relevant to policies on trial is almost twice as large if the total number of participating localities is small (see Appendix Table A.7). This suggests that local politicians may internalize the higher political reward they may receive when they participate in successful experiments with fewer competing politicians.[25]

To examine the dynamic patterns of fiscal inputs associated with policy experimentation, we trace domain-specific expenditure around the time of each county-domain's first engagement in policy experimentation during our sample period. Appendix Figure A.12 plots the yearly estimates five years before and four years after the start of the experiment (four years is the average duration of experiments). We observe little evidence of a pre-trend in domain-specific fiscal expenditure leading up to the first policy experiment. Right after being assigned a policy experiment, local politicians begin to spend significantly more in the corresponding policy domain.

One may be concerned that the increased local fiscal expenditure, rather than reflecting local government's political incentives and efforts, is substituting for the lack of central government's fiscal support for the specific experiment.[26] We find this unlikely. First, the increase in domain-specific fiscal expenditure during policy experiments is observed even if the experimentation guideline explicitly provides fiscal support from the central government (see Appendix Table A.10). Second, we conduct the regression analysis at policy-county level instead, controlling for experiment FEs, and thus exploiting variations in political incentives within experiment across participating localities. We observe a consistent pattern: local politicians that have stronger career incentives are spending more fiscal resources during the experiment compared to other politicians participating in the same experiment (see Appendix Table A.11).

Fiscal expenditure outside of experimentation Importantly, such experimentation-induced additional fiscal expenditure may not be sustained when a policy becomes national. Indeed, we do *not* find fiscal expenditure increasing in corresponding domains among non-

term/position, position and official rank in the bureaucratic system. See Appendix B.1 for details.

24. We find similar results zooming into politicians > 50 years old and exploiting the sharp drop in promotion eligibility after 58 years old; which suggests that politicians' incentives likely play a causal role in generating the observed fiscal responses. The results are presented in Appendix Table A.6.

25. Interestingly, the increased fiscal expenditure in the experimentation domain is stronger if the locality is engaged in one experiment (the effect of each experiment increases by 50%, see Appendix Table A.8), reflecting a multi-tasking problem faced by the local politicians.

26. Interestingly, local politicians in richer jurisdictions do *not* show stronger fiscal responses to policy experiments (see Appendix Table A.9).

experimentation sites when the same policy rolls out to the entire country. This is the case regardless of the career incentives of the local politicians at these non-experimentation sites (see Table 2, Panel B).[27] Moreover, among experimentation sites, increase in fiscal expenditure on the experimentation domain stops after the completion of experimentation (see Appendix Table A.12). Again, this indicates that the local politicians' heightened efforts are specifically targeted toward the experiment itself.

Other dimension of efforts during experimentation Beyond increased domain-specific fiscal expenditure during experimentation, we find that local politicians also exert efforts to differentiate in their implementation of the experimental policies. Differentiation can signal effort and potentially 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. We construct pairwise text similarity among documents issued by local governments on the corresponding policy experiment, calculated using Latent Similarity Analysis (LSA). We observe that, when local politicians have strong career incentives, they tend to differentiate more than their colleagues in terms of implementation details, reflecting an increase in local politicians' efforts to stand out in achieving good results in the experiment. Appendix G presents the details of the empirical specification and discussion of the results.

7 Is the central government sophisticated in interpreting experimentation outcomes?

In this section, we ask whether the experimentation outcomes are interpreted in a sophisticated manner by the central government. Specifically, we focus on exogenous shocks that affect experimentation outcomes but are fundamentally *unrelated* to the experimental policies themselves, and therefore should *not* be taken into account when evaluating policy effectiveness. We examine whether such shocks influence how the policies on trial are assessed for national roll-out, with Section 7.1 focusing on locality-specific shocks and Section 7.2 focusing on politician-specific shocks.

To the extent that these shocks affect national policy decisions, it reflects a lack of sophistication of the central government when interpreting experimentation outcomes,

27. 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 (Fang, Liu, and Zhou 2020).

regardless of its objective function. In the language of the framework presented in Section 4, we test whether $\delta_2 \neq 1$.

7.1 Experimentation outcomes and locality-specific shocks

When evaluating experimentation outcomes, is the central government able to disregard locality-specific shocks that may impact observed experimentation outcomes but are orthogonal to the underlying policy effectiveness? In particular, does a local fiscal windfall during experimentation, which may substantially improve local socioeconomic outcomes but is unrelated to the innate effectiveness of the trial policy, increase the likelihood that the central government decides that the policy is successful?

We focus on land revenue (i.e., land conveyance fees) received by the county governments for converting agricultural land for residential use during the period of experimentation. Land conveyance fees are by far the most important source of local fiscal revenue, accounting for more than 75% of total budgetary income in recent decades (Lan 2021). Local land revenue is transparently reported and visible to the central government.[28] We follow Chen and Kung (2016) and use the *ratio of land suitable for construction* $\times$ *national interest rate* to instrument for each county's land revenue windfall in a given year (conditional on county and year fixed effects). When conveying rural land for residential use, the Chinese government enforces an architectural safety standard that considers land with a slope of 15 degrees or less to be safe for real estate construction. Thus, different counties have different land conveyance potentials based on terrain features, and are differentially affected when there is a macroeconomic demand shock in the real estate market, such as a change in the national interest rate. Since the initial stock of land type is pre-determined, and the national interest rate is unlikely to be influenced by an individual county, changes in land revenue induced by the interaction of these two factors are likely exogenous to other county-level outcomes.

We evaluate whether land revenue fluctuation caused by the interaction of these two factors during policy experimentation among experimentation site — which are unrelated to the experimentation and policy effectiveness *per se* — may affect the chance that the trial policy gets rolled out to the entire country. We estimate the following two-stage least-squares specification:

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

28. Since 2002, every land auction (the method by which the local governments generate land revenue) is required by law to be publicized on a central government website.

where $Land\_revenue_{ipt}$ is the log level of land conversion revenue obtained by county i, serving as an experimentation site for policy p, in year t. The instrumental variable is the interaction term between the geographic constraint on experimentation site i's land supply (determined by its land slope) and 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.[29]

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 (First stage f-stat=622.9, see Appendix Table A.14). Table 3, 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.[30] In other words, when policy experimentation is conducted in localities that coincidentally experience temporal shocks that could improve the policy outcome, the central government does *not* fully discount these factors, but instead at least partially attributes the policy outcomes to the underlying policy effectiveness. This results in biased policy learning and policy choices. Interestingly, the central government's roll-out decisions are more affected by land revenue windfalls in experiments with fewer participating sites (see Appendix Table A.16), consistent with the fact that each locality plays a bigger role in shaping the experimentation outcomes that the central government observes.

We implement several additional tests to assess the validity of our empirical strategy. First, we estimate how the leads of the IV affect land revenue. Future national interest rate changes should not affect the land revenue in the current period (apart from short-run auto-correlation in interest rates). As shown in Appendix Figure A.13, we indeed observe that, while a contemporaneous credit shock in year t has a very large and significant impact on a county's land revenue in the same year, future shocks in $t+1$ and $t+2$ both have minimal impacts on current land revenue.

Second, we examine whether a placebo IV — *ratio of land between 15 and 30 degrees* $\times$

29. Following Chen and Kung (2016), we also control for characteristics at the county level (log population and lagged local GDP growth rate), politician level (age, educational attainment, whether they are a member of the Youth League, previous prefectural government experience, whether they share a birth-county connection with the prefectural leader, and current year in office). Excluding these control variables has minimal impacts on our IV estimates.

30. Our baseline analysis is conducted at the county-policy-year level – if a county has two ongoing policy experiments in a given year, it shows up as two county-policy units in our data in that year. Alternatively, we can conduct the analysis at the county-year level, in which case the outcome of interest becomes "how many policy experiments in that county in that year turned into national policies," rather than "did the policy experiment in that county in that year turn into a national policy." As shown in Appendix Table A.15, our findings are robust to this alternative way of structuring the data.

contemporaneous national interest rate — affects local land revenue. Since the government's official cutoff for real estate development is 15 degrees, only land plots with slopes below this cutoff would matter for real estate construction. As shown in Appendix Table A.17, Panel B, this is indeed the case: land plots between 15 and 30 degrees contribute little to local land revenue (F=0.2).

Third, we conduct a falsification test of the second stage analysis, replacing the instrumented land revenue during the experimentation with instrumented land revenue that occurs after the experimentation ends. Specifically, we use the *ratio of land suitable for construction* $\times$ *national interest rate at* $t+5$ as the instrument for land revenue at $t+5$ (since the vast majority of policy experiments conclude within five years). As we can see in Appendix Table A.17, Panel C, while there is a very strong first stage, land revenue at $t+5$ has a precisely estimated null effect on the roll-out of policies being experimented in year t.

7.2 Experimentation outcomes and politician-specific shocks

When evaluating experimentation outcomes, does the central government exclude politician-specific shocks that may impact observed experimentation outcomes but are orthogonal to the underlying policy effectiveness? In particular, we examine whether changes in local politicians' career incentives (and thus changes in effort, as shown in Section 6) due to local politicians' routine turnover affects the central government's policy learning and increases the likelihood of the trial policy being evaluated as successful.

We focus on local politicians' turnover taking place among experimentation sites *after* the beginning of policy experimentation in the local region. This allows us to isolate changes in local politicians' career incentives caused by the turnover that are unrelated to either the underlying effectiveness of the trial policy or the local government's (potentially endogenous) initial participation in the policy experiment. Specifically, we estimate the following model:

$$\begin{aligned} y_p = \alpha \cdot Turnover_{ip} + \beta_1 \cdot Turnover_{ip} \times IncreaseIncentive_{ip} \\ + \beta_2 \cdot Turnover_{ip} \times DecreaseIncentive_{ip} + \gamma_t + \delta_m + \theta_n + \varepsilon_{ipmnt}, \end{aligned}$$

where y_p is the indicator of experiment p being evaluated as successful and rolled out to the entire country; and $Turnover_{ip}$ is the indicator of a change in the Party Secretary of prefecture i during the experimentation period of policy p among experimentation sites. A change in $Incentive_{ip}$ is calculated based on the difference in career incentives between the incumbent at the beginning of the experiment and that of his or her immediate successor (the baseline career incentives measure, following Wang, Zhang, and Zhou (2020),

is described in Appendix B.1). We separate local political turnovers that result in either an increase ($IncreaseIncentive_{ip}$) or a decrease ($DecreaseIncentive_{ip}$) in the politicians' career incentives. We include a full set of year fixed effects (γ_t), ministry fixed effects (δ_m), and province fixed effects (θ_n), allowing us to isolate the effects due to the (asynchronous) rotation of local politicians.

Table 3, Panel B, presents the results. We observe that local politician rotation *after* the start of the experiment does not affect the likelihood of the trial policy's national roll out. However, when the incoming politician has stronger upward career mobility potential than the outgoing politician (i.e, younger versus retiring), the trial policy becomes substantially more likely to be assessed as successful and rolled out nationwide. The opposite pattern is observed when local politician rotation results in a reduction in politicians' promotion prospects and career incentives. This suggests that, when policy experiments are conducted in localities that experience politician-related shocks that could improve the experimentation outcome, the central government incorrectly attributes the outcome at least partially to policy effectiveness, again resulting in biased policy learning. Consistent with previous findings, the impact of political rotations on roll-out decisions is more pronounced among small-scale experiments (see Appendix Table A.18).

The impact of changes in political incentives due to politicians' rotation during the experimentation period is robust to alternative measures of political incentives — in particular, if we examine the sharp changes in promotion incentives among politicians above or below the 58 years-old cut-off (see Appendix Table A.19, Panel A). Such impact is observed even among relatively low-stakes policies not appearing in the national Five Year Plans (see Panel B).[31] Moreover, our findings are unlikely to be driven by a spurious correlation between political rotation and policy experimentation success. First, such an impact of changes in political incentives is consistently observed even if we focus only on the political rotations toward the end of the experimentation period (see Panel C). When political rotation occurs toward the end of the experimentation period, the incoming politicians' ability to directly affect the experimentation outcomes becomes limited, although they can influence local economic performance during certain years as a result of the changes in career incentives. Second, reassuringly, we do *not* observe similar effects with the rotation of politicians that happened either before the start of the policy experimentation or at least five years after the beginning of the experimentation period (see Panels D.1 and D.2, respectively).

31. The rotation of local leaders is decided by the Organization Department, rather than by the ministries that are in charge of most policy experiments. Thus, it is unlikely that such rotations are catered to specific policies on trial, especially for the low-stakes policies.

8 What are the implications on policy learning and policy outcomes?

Having documented three facts about China's policy experimentation in Sections 5-7, we now examine their implications for the central government's policy learning and the effectiveness of national policies originating from such experimentation. In Section 8.1, we discuss such implications under the assumption that the central government of China aims to learn about policy's average treatment effects. In Section 8.2, we discuss alternative objectives of policy experimentation beyond learning about policies' average treatment effects.

8.1 If experimentation objective is to learn about policies' ATE

8.1.1 Experimentation outcomes and policies' national roll-out

While we do not directly observe how the central government evaluates policy experiments and decides on policies' roll-out, we can infer the decision rule by examining which estimators of experimentation outcomes most strongly predict the corresponding policies' national roll-out.

We begin with a simple estimator of experimentation effects that compares experimentation sites' local economic performance before and after the experiments, averaged across all experimentation sites. Figure 4, Panel A, presents the correlation between the estimated experimentation effects (on the horizontal axis) and the decision to roll out a policy nationally (on the vertical axis). There is a strongly positive correlation between the two: a one standard deviation increase in experimentation outcomes, measured as the average differences in local economic performance (GDP per capita) before and after the experiments, is associated with a 7.0 percentage point (or 17.9%) higher likelihood of the experimental policy turning into a national policy. This correlation is largely unchanged if we control for experimentation year fixed effects or ministry fixed effects — thus simultaneously comparing policies that have been evaluated by the same minister (see Appendix Table A.20, Panel A).[32]

32. In Appendix Table A.21, we re-estimate the correlation between the estimated experimentation effects and the national roll-out decision, winsorizing the sample by 2.5 percent at either the top or bottom end of the experimentation effects distribution. The baseline pattern is unchanged when we drop experiments beyond the top 2.5 percentile of experimentation effects. The baseline pattern remains, although it becomes weaker, if we drop those at the bottom 2.5 percentile of the experimentation effects. This is consistent with policies generating bad outcomes being disproportionately salient to the central government, which we discuss in greater detail in Section 8.2.1. The sample for this analysis starts in 1993, which was when

The comparison of experimentation sites' economic performance before and after the experiments is not informative of the experimental policy's average treatment effect, since it does not account for positive site selection and non-representative experimental situations. By controlling for province-specific time trends, one can partially control for the differential growth trajectories that the experimentation sites might be experiencing before the experiments start. One could also use synthetic control, following methods such as Xu (2017), to match experimentation sites with a weighted sample of non-experimentation sites based on five-year pre-experimentation trends in local socioeconomic conditions. The correlations between these estimated experimentation effects and policies' national roll-out are plotted in Figure 4, Panels B and C, respectively (and in Appendix Table A.20, Panels B and C, in regression form). These more sophisticated estimates of experimentation effects, which would have been more informative of policies' average treatment effects, no longer predict whether policies roll out nationwide.

Assessing the magnitude Leveraging the estimates from Sections 5-7, we perform a back-of-the-envelope calculation on how much the national policy roll-out would be affected by the presence of positive site selection, local governments' strategic efforts, and the central government's naivety in interpreting the experimentation outcomes (see Appendix H for details).[33] A 1% increase in fiscal revenue for all experimentation sites would increase the corresponding policy's national roll-out probability by 1.8 percentage points. Linking this number to the average (pre-experimentation) difference in fiscal revenue between experimentation and non-experimentation sites (20.5%), we calculate that positive site selection inflates the national roll-out rate of an average policy experiment by 36.9%. Linking this number to the strategic (and extra) fiscal expenditure induced by policy experimentation (8.1%), we calculate that non-representative fiscal resources inflate the national roll-out rate of an average policy experiment by an additional 24.1%.[34]

county-level socio-economic data started to be reliably published in the Statistical Yearbooks.

33. It is important to note that one cannot easily decompose the separate roles of positive site selection and endogenous local efforts. 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.

34. In addition, according to our estimates in Section 7.1, a 1% increase in local politicians' promotion incentives would increase the corresponding policy's national roll-out probability by 0.68 percentage points. Linking this elasticity to the average difference in local politicians' incentives between experimentation and non-experimentation sites (1.3%), we calculate that non-representative political incentives inflate the

8.1.2 National policy outcomes

Positive experimentation site selection and extra efforts among local politicians during the experiment both could result in better experimentation outcomes. If the central government does not take these factors into account when they select experimental policies to roll out, then one would expect policy outcomes during experimentation to be considerably better than the national outcomes when the policies are rolled out to the entire country.

Throughout this sub-section, when constructing measures of policy outcomes during either experimentation or roll-out, we focus on the subset of policies that are in the economic domain. This allows us to proxy policy outcomes using local economic indicators such as economic growth.[35]

Do national outcomes shrink in comparison to experimentation outcomes? We begin by examining the potential shrinkage of experimentation effects across all experimental economic policies. In Figure 5, Panel A, for each of the economic policies that have been tried and then rolled out to the entire country, we plot the experimentation effects on local economic growth (estimated as the differences of economic performance among experimentation sites before and after the experiments) against the national effects (estimated as the differences of economic performance among non-experimentation sites before and after the corresponding policies roll out to the country). Panel B plots the distribution of the differences of experimentation and national policy effects. One observes that many policies (71.1%) fall below the 45 degree line, reflecting smaller effects during the national roll-out.[36] In fact, while the (naively estimated) experimentation effects strongly predict policies' national roll-out, they do *not* predict the corresponding policies' national average effects.[37]

Such shrinkage is unlikely to be driven by local politicians' exaggerated reporting of local economic performance during experiments. We find similar patterns of shrinkage if we: *(i)* instead focusing on policy effects on local fiscal revenue, an indicator of local

national roll-out rate of an average policy experiment by an additional 2.2%.

35. Two-thirds of all policy experiments are related to economic policies according to our definition. Our findings are robust to different characterizations of economic policies.

36. We replicate Figure 5, controlling for the number of experimentation sites (namely, the sample size for each experiment). The results are presented in Appendix Figure A.14. This does not qualitatively or quantitatively change the baseline pattern of shrinkage.

37. We plot the regression coefficients that use experimentation effects to predict various estimates of the policies' national effects in Appendix Figure A.15. We observe that the experimentation effects are moderately predictive of the national average effects (equally weighted across all localities); the regression coefficients $= 0.03$. As the weights placed on non-experimentation sites increase, the experimentation effects become substantially less predictive of the national policy effects.

economic performance that is unlikely to be manipulated by the local politicians (see Appendix Figure A.16 and Figure A.17, Panel A); and *(ii)* correct for local economic growth (mis)reporting using local luminosity from satellite images, following Martinez (2022) (see Panel B).

The shrinkage of experimentation effects as a policy rolls out could result from a combination of the lack of representativeness of experimentation sites (both in terms of socioeconomic characteristics and local politicians' effort) and the central government's naive inference.[38] To gauge the relative importance of these factors, we regress the policy effects' shrinkage on the gap between the experimentation effects estimated using naive, simple mean difference and synthetic control, where site selection and endogenous efforts may be taken into account. As shown in Appendix Figure A.19, the gap between the naive estimator and the synthetic control estimator is positively correlated with the extent to which experiment effects deflate. This suggests that the deflation in effect sizes is not merely a result of regression to the mean, and could have been partially mitigated if the government had been more sophisticated in their interpretation of the experimentation outcomes.

The shrinkage of experimentation effects is best illustrated in the context of a specific policy experiment on local fiscal empowerment. In order to foster economic growth, the central government initiated an experiment that provides more fiscal autonomy to the counties participating in the experiment (see Appendix A.2 for policy details). Between 2003 and 2013, more than 1,100 counties were selected as experimentation sites. The experimentation sites were positively selected during the first half of the experiment and moved to negative selection toward the end of the experiment (t-stat$>$ 10 in 2004, t-stat<-3 in 2017, see Appendix Figure A.20 for more details). We use a staggered event study design to estimate the treatment effects of the introduction of such policy experiment on local economic performance (controlling for county and year fixed effects), and we separately report the coefficients among the subsamples of experimentation counties in the early rounds (positively selected) and the later rounds (negatively selected).

We find that counties that had higher pre-experimentation GDP per capita benefited from the experiment, while the poorer counties experienced worse subsequent local economic development (see Appendix Figure A.21).[39] The local fiscal empowerment exper-

38. Differences between experimentation and national policy outcomes could be driven by general equilibrium effects. Whether general equilibrium mechanisms lead to reduced or larger effects is often theoretically ambiguous (e.g., Muralidharan and Niehaus (2017)). Interestingly, we do *not* observe less reduction of experimentation effects among experiments that are aimed at improving short-run outcomes (see Appendix Figure A.18).

39. Such patterns of heterogeneity by pre-experimentation local economic conditions do *not* merely reflect a general equilibrium effect or an early-mover advantage in reform. Less-developed counties participating

iment did not lead to a national policy, likely because of the negative selection in experimentation sites in the latter stage of the experiment. Had the policy been rolled out to the entire country, it would likely have generated a net zero effect, with both winners and losers (see Appendix Figure A.23).

Do regions similar to experimentation sites benefit more? When experimental policies roll out to the entire country, localities similar to experimentation sites may benefit more from the new policy. To examine this hypothesis, for each experiment that eventually leads to a national policy, 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 socioeconomic conditions (local GDP per capita, local fiscal income, and fiscal expenditure), as well as the local officials' career incentives. We then examine, among localities that did not participate in an experiment, whether the corresponding national policy leads to faster local economic growth when a specific county is similar to the experimentation sites for that policy.

We estimate the following specification, identifying differential policy effects on a specific locality as a result of the composition of the experimentation sites where the policy was originated from:

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

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 4. Panel A shows the results when we calculate M_{cp} based on the vector of socioeconomic conditions; Panel B shows those based on local officials' career incentives. We observe that, when an experimental policy rolls out to the entire country, localities that did not participate in an experiment but are socioeconomically similar to the experimentation sites benefit significantly *more*. Moreover, non-experimentation sites with local politicians facing similar career incentives as the experimentation sites are also better off when the trial policies roll out nationwide. These results are robust to different indices chosen to compute the distance (See Appendix Table A.22).

These results suggest two things. First, policies originating from unrepresentative experiments differentially benefit some regions over others, depending on the sample composition of the experimentation sites. Second, experimentation may structurally allow for better tailoring of policies to benefit from greater efforts by local officials. Given that the

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.22).

experimentation sites are overwhelmingly positively selected in terms of local political and economic conditions, this would generate distributional consequences: positive selection of sites may produce a portfolio of policies that systematically favor regions with better socioeconomic conditions and more incentivized politicians at the expense of their less-developed and less incentivized counterparts, thus leading to greater inter-regional inequality throughout China.

8.2 Alternative experimentation objectives

In Section 8.1, we evaluated the implications of the structure of policy experimentation for policy learning and policy outcomes, assuming that the central government aims to learn about the average treatment effects of the policies. We ultimately do not observe the central government's objective function, and in this section, we discuss alternative objectives of policy experimentation that are both related and unrelated to learning.

8.2.1 More complex learning-related objectives

Learning about tail risks In addition to (or instead of) learning about the average effects of policies, policy experimentation might be critical for the central government to assess the potential risks associated with the policy on trial. To examine this possibility, for each policy experiment, we count the number of experimentation sites that fall below a certain percentile across all localities in the nation in terms of local GDP growth during the period of experimentation, and investigate whether this measure is predictive of the national roll-out of the corresponding experimental policy. Appendix Figure A.24 presents the estimated coefficients across the percentile thresholds, which range from 0 to 50th percentile. The presence of experimentation sites that fell below the 10th percentile of local GDP growth nationwide substantially decreases the chance that the policies roll out to the country, and this remains true even after controlling for the policies' underlying average treatment effects (estimated based on before and after differences in GDP growth during the experimentation stage). While it is not obvious that one could attribute the low growth performance to the policy experiment, this result suggests that the central government may be particularly sensitive to those instances when they evaluate the experimentation outcomes.

Incorporating decision-makers' subjective expected utility In addition to learning about the true underlying treatment effects, the central government may hold subjective expected utility when designing the policy experimentation. This may justify unrepresen-

tative choices of experimentation sites. To evaluate the importance of subjective expected utility, we conduct quantitative exercises following Banerjee et al. (2020). We simulate the optimal experimentation design, parameterizing the model based on data from the 25th, 50th, and 75th percentile of Chinese policy experiments in terms of their degree of positive selection. Appendix E.1 provides details of the simulation procedure.

We find that, when the central government places greater 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, less than 5% of the optimal designs for these experiments would induce positive selection with *t-statistics* > 1, with the optimal *t-statistic* never exceeding 2.6 — substantially lower than the positive selection that actually occurs.[40]

Incorporating experimentation sites' welfare The central government may incorporate considerations about the welfare of the experimentation sites, following Narita (2021). In particular, deviation from full randomization may be justified in an optimal experimentation design when the sample size is small, and there exist sufficiently large heterogeneous treatment effects as well as heterogeneous welfare from receiving the treatment.[41] We again simulate the optimal experimentation design, parameterizing the model based on China's policy experimentation setup; Appendix E.2 provides details of the simulation. We find that the central government would have to place almost the entirety of its welfare weight on the locations that were selected as the experimentation sites in the early waves in order to justify the observed degree of positive selection. In other words, the observed level of positive selection could be optimal only if extreme *ex ante* inequality is inherent to the central government's objective function.

8.2.2 Politically motivated objectives

Political patronage 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 shape the selection of experimentation sites. This could be due to exchange of favors, higher trust among political patrons, and ministers' stronger control over local implementation.

40. We additionally test two extensions on the model presented: *(i)* we allow for the quality of experimental information (or equivalently, policy execution) to vary with the local county's GDP; and *(ii)* we allow counties to opt into 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.

41. This can be captured as either experimentation subjects' willingness to pay, or benevolent social planners' welfare weights across subjects.

We define a province as connected to a ministry if the current minister used to work full-time in that province before assuming his current position, following Jia, Kudamatsu, and Seim (2015). To investigate the role of political patronage in the selection of experimentation sites, we exploit the inter-temporal changes in a region's connection to each ministry caused by the turnover of central ministers. We estimate the number of experiments assigned to province p by ministry m in year t, as a function of whether the minister of ministry m in year t used to work full-time in province p (controlling for year fixed effects and province-by-ministry fixed effects). 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. In Appendix Figure A.25, we plot the event study estimates around ministers' turnover; reassuringly, we do not observe a pre-trend.

As shown in Appendix Table A.23, Panel A, as soon as a locality becomes connected to a minister, the number of experiments assigned to that region increases by 28.8%. 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 site selection was done via voluntary participation (see Panels B and C).

Demand for political stability In addition to learning about policies' impact on local economic performance, the central government may be concerned with maintaining political stability during socioeconomic reforms. To evaluate this possibility, we first examine whether social and political unrest in a particular prefecture is correlated with its chance of being selected as an experimentation site. We exploit within-region, across-time variations in occurrences of unrest: estimating whether prefecture p engages in policy experimentation in year t as a function of unrest occurred in prefecture p during the previous year $t-1$ (measured as unrest event counts in GDELT, following Beraja et al. (2023)), controlling for prefecture and year fixed effects. We find a robust pattern that prefectures that have experienced social and political unrest in the preceding year are significantly and substantially less likely to become experimentation sites (see Appendix Table A.24). This suggests that an unstable local environment could be a veto condition that precludes participation in policy experimentation.

Next, we investigate whether occurrence of social and political unrest during experimentation affects the likelihood of experimental policies' national roll-out. Specifically, we run a policy-prefecture level regression, regressing whether the experimental policy is rolled out to the entire nation on the number of unrest events in the corresponding prefecture when the experiment starts, controlling for prefecture and year fixed effects. We find

that, conditional on observed experimentation outcomes on local economic performance, as measured as in Section 8.1.1, unrest episodes are associated with substantially lower chances that the local experimental policies would eventually become national policies (see Appendix Table A.25, Panel A). This suggests that avoidance of policy disruptions that are associated with social and political unrest may be a salient criterion when the central government evaluates experimentation outcomes. This could at times contradict its objective of selecting policies that maximize economic performance. To further establish the causal effects of social and political unrest on policy experimentation adoption and roll-out, we follow Beraja et al. (2023) and use local weather conditions to instrument for protest occurrence.[42] As shown in Appendix Tables A.25, Panel B, weather-induced variations in protests strongly predict the national roll-out of experimental policies. Since weather-induced variation in protest occurrence is orthogonal to experimentation itself, this result indicates another potential error in attribution: any protest, regardless of whether they are caused by the experiment, could stop the policy's national roll-out.

9 Conclusion

In this project, we examine China's extensive policy experimentation over the past four decades, one of the largest undertakings of systematic policy learning in recent history. We document three facts about China's policy experimentation. First, policy experimentation sites are positively selected for characteristics such as local socioeconomic conditions. Second, the experimental situation during policy experimentation is unrepresentative: local politicians exert strategic efforts and allocate more resources during experimentation, which may exaggerate policy effectiveness. Third, the central government is not fully sophisticated when interpreting experimentation outcomes, indicating that the positive sample selection and unrepresentative experimental situations might not be fully taken into account for national policy choices. These facts imply that China's unrepresentative policy experimentation could lead to biases in policy choices and shrinkage in policy effectiveness during national roll-out, if the central government intends to learn about the policy's average effects in a representative locality with representative local politician's incentives.

We highlight that policy learning and policy experimentation inevitably take place in complex environments with various constraints and distortions. The political and bu-

42. Since this empirical strategy relies on the detailed timing of each protest, which is only available after 2014, the sample size becomes substantially smaller for this exercise.

reaucratic environment could affect the initiation of policy experimentation, its structure and implementation, and bias in the information gathered 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).[43] Rather than the informational free-riding and under-experimentation observed in federalist systems, political centralization — in a context such as China where local government officials compete and differentiate their implementation activities in order to increase their chances of promotion —could overcome tendencies of under-experimentation.[44]

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.[45] Our findings that policies originating from unrepresentative experimentation could disproportionately benefit richer regions demonstrate yet another manifestation of regulatory capture — in this case, 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).[46] Moreover, our findings point to a fundamental trade-off that the central government faces: structuring political incentives in order to stimulate politicians' efforts to improve policy outcomes, while making sure that such incentives are not exaggerated

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

44. Following List (2020), Goldszmidt et al. (2020), and Holz et al. (2020), we examine the SANS conditions of our study to shed light on the extent to which our findings may apply to other institutional contexts. On *selection*, the sample of our study is the universe of policy experiments conducted in China over the past four decades. On *attrition*, all announced policy experiments are included in the sample. On *naturalness*, the type of policy experimentation that we study has been the key institution for policy making in China for forty years, and most high-stakes policy ideas have to be tested this way before becoming national policies. On *scaling*, since our findings concentrate on cases where the central government leads policy experimentation in a top-down manner, we think the key "non-negotiable" for external validity is that the central government plays the leading role in conducting policy experiments; this would be relevant in many contexts as governments, especially those in the developing countries, are explicitly learning from China's policy experimentation (e.g., Vietnam).

45. As China develops and low-hanging fruit for policy improvements diminishes, it may become increasingly important to carefully structure policy experimentation in order to achieve better policy-making.

46. 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).

during the experimentation phase, so that policy learning remains unbiased. Dynamic experimentation could be a solution (e.g., Kasy and Sautmann 2021). More generally, future work on mechanism designs that could improve the efficiency of policy learning could be of great academic importance and policy relevance.

Our work does *not* address the overall benefits (or costs) of experimentation relative to a counterfactual of no experimentation at all. This study does not examine, for example, which policies are subject to experimentation in the first place, and which major policy disasters may have been avoided because of the experimentation. Evaluating the overall policy-making cycle would be a fascinating, important, and challenging undertaking that we leave for future work.

References

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.

Al-Ubaydli, Omar, John A List, Danielle LoRe, and Dana Suskind. 2017. "Scaling for economists: Lessons from the non-adherence problem in the medical literature." *Journal of Economic Perspectives* 31 (4): 125–44.

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.*

Al-Ubaydli, Omar, Min Sok Lee, John A List, Claire L Mackevicius, and Dana Suskind. 2021. "How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling." *Behavioural public policy* 5 (1): 2–49.

Bai, Chong-En, Chang-Tai Hsieh, and Zheng Song. 2020. "Special deals with chinese characteristics." *NBER Macroeconomics Annual* 34 (1): 341–379.

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.

Beraja, Martin, Andrew Kao, David Y Yang, and Noam Yuchtman. 2023. "AI-tocracy." *The Quarterly Journal of Economics* 138 (3): 1349–1402.

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.

Blanchard, Olivier, and Andrei Shleifer. 2001. "Federalism with and without political centralization: China versus Russia." *IMF staff papers* 48 (1): 171–179.

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. 2022. "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 China's 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.

Goldszmidt, Ariel, John A List, Robert D Metcalfe, Ian Muir, V Kerry Smith, and Jenny Wang. 2020. *The value of time in the United States: Estimates from nationwide natural field experiments.* Technical report. National Bureau of Economic Research.

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

He, Guojun, Shaoda Wang, and Bing Zhang. 2020. "Watering down environmental regulation in China." *The Quarterly Journal of Economics* 135 (4): 2135–2185.

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 China's 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. 2021. "How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities." *American Economic Review* 111, no. 5 (May): 1442–80.

Holz, Justin E, John A List, Alejandro Zentner, Marvin Cardoza, and Joaquin Zentner. 2020. *The* 100*millionnudge* : *Increasingtaxcomplianceofbusinessesandtheself* − *employedusinganatu* Technical report. National Bureau of Economic Research.

Jia, Junxue, Qingwang Guo, and Jing Zhang. 2014. "Fiscal decentralization and local expenditure policy in China." *China Economic Review* 28:107–122.

Jia, Ruixue, Masayuki Kudamatsu, and David Seim. 2015. "Political Selection in China: The Complementary Roles of Connections and Performance." *Journal of the European Economic Association* 13, no. 4 (August): 631–668.

Jiang, Junyan. 2018. "Making bureaucracy work: Patronage networks, performance incentives, and economic development in China." *American Journal of Political Science* 62 (4): 982–999.

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

Kasy, Maximilian, and Anja Sautmann. 2021. "Adaptive Treatment Assignment in Experiments for Policy Choice." *Econometrica* 89 (1): 113–132.

Kornai, Janos. 1959. *Overcentralization in economic administration: A critical analysis based on experience in Hungarian light industry.* London, Oxford UP.

Lan, Xiaohuan. 2021. *Embedded Power: Chinese Government and Economic Development.*

Li, Hongbin, and Li-An Zhou. 2005. "Political turnover and economic performance: the incentive role of personnel control in China." *Journal of public economics* 89 (9-10): 1743–1762.

List, John A. 2020. *Non est disputandum de generalizability? A glimpse into the external validity trial.* Technical report. National Bureau of Economic Research.

———. 2022. *The voltage effect: How to make good ideas great and great ideas scale.* Currency.

Martinez, Luis R. 2022. "How much should we trust the dictator's GDP growth estimates?" *Journal of Political Economy* 130 (10): 2731–2769.

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.

Muralidharan, Karthik, and Paul Niehaus. 2017. "Experimentation at scale." *Journal of Economic Perspectives* 31 (4): 103–24.

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

Narita, Yusuke. 2021. "Incorporating ethics and welfare into randomized experiments." *Proceedings of the National Academy of Sciences* 118 (1).

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.*

Simonsohn, Uri, Joseph P Simmons, and Leif D Nelson. 2020. "Specification curve analysis." *Nature Human Behaviour* 4 (11): 1208–1214.

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.

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.*

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.

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

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

Figures and tables

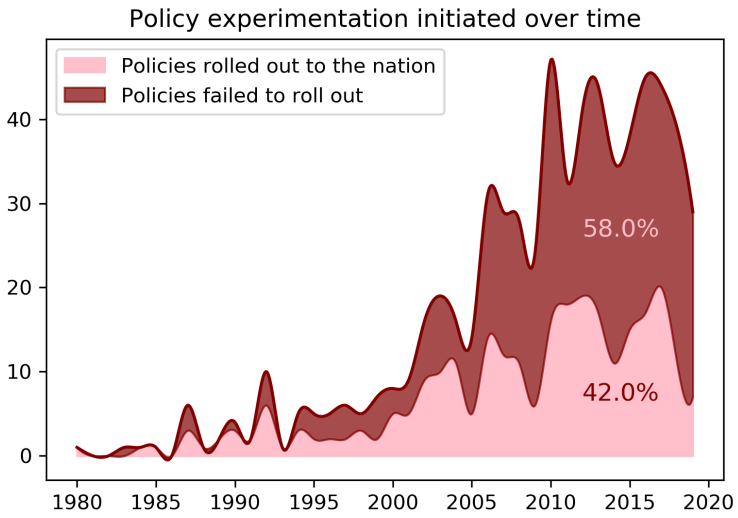
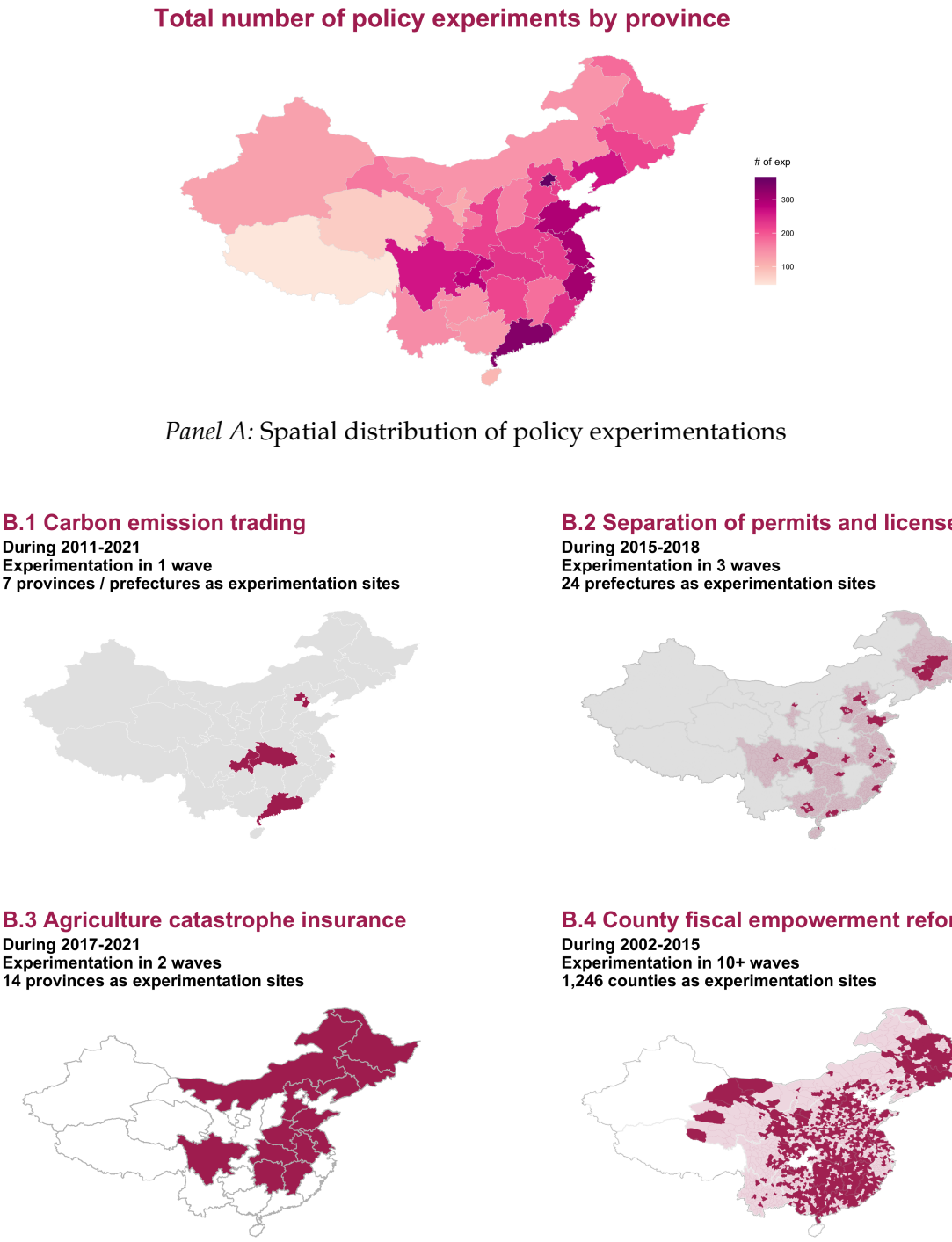


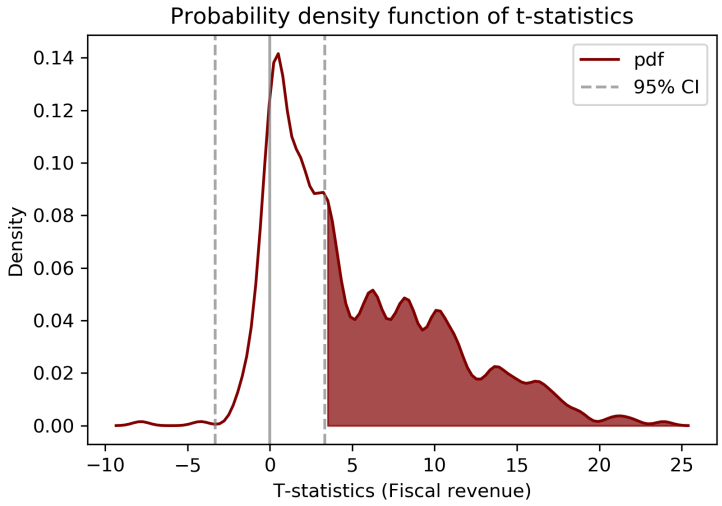
Figure 1: This figure plots the number of policy experiments initiated over time. The share of successful experiments that eventually rolled out to the entire country is indicated by the area shaded in pink; the share of unsuccessful policies that failed to roll out to the entire country is indicated by the area shaded in red.



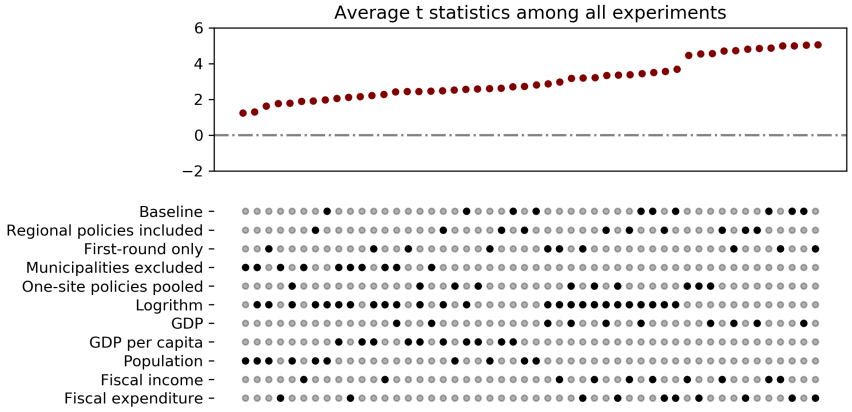
Panel A: Spatial distribution of policy experimentations

Panel B: Examples of policy experimentation

Figure 2: Panel A counts the total number of policy experiments that each province has been involved in between 1980 and 2020 (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 gray indicate parts of the country that eventually received the policies when they rolled out. 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.

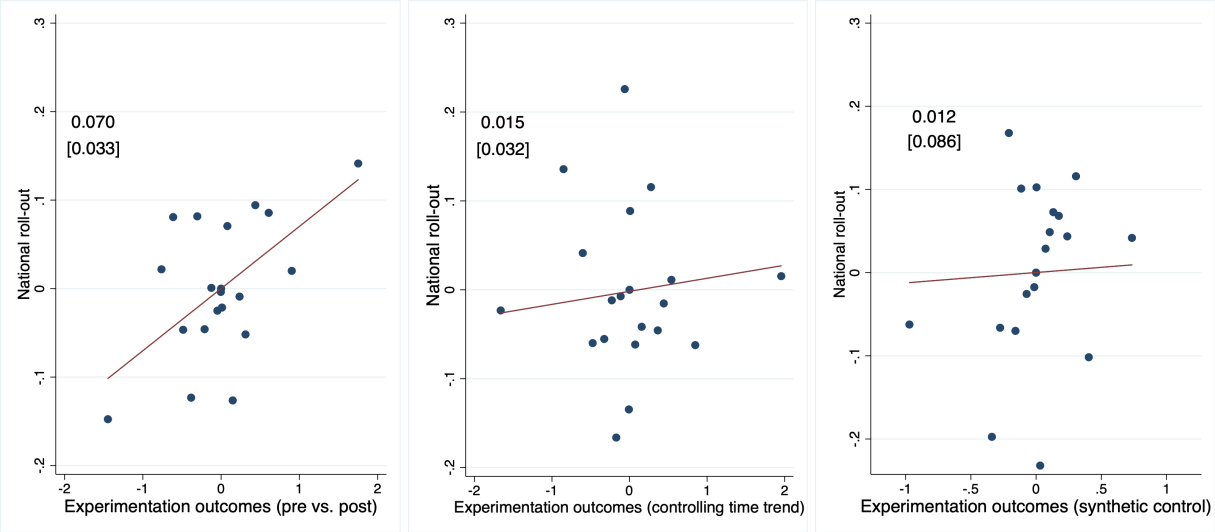


Panel A



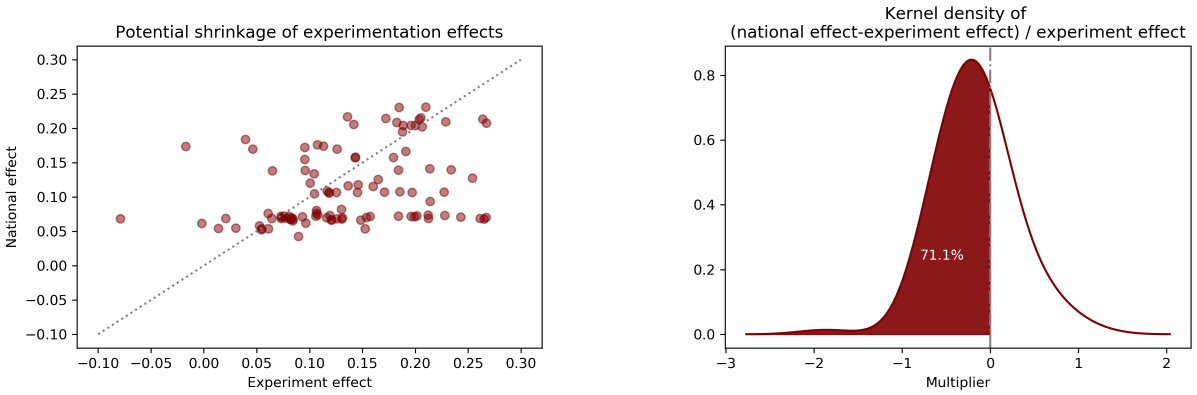
Panel B

Figure 3: Panel A plots the distribution of t-statistics from the representativeness test for experimentation sites, calculated based on fiscal expenditure. 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. Panel B summarizes the t-test results for other specifications. We presents those results as a specification curve, à la Simonsohn, Simmons, and Nelson (2020), with solid circles indicating the combination of plausible specifications.



A. Pre vs. post B. Controlling for trend C. Synthetic control

Figure 4: These plots visualize the correlations between policy roll-out rate and different measures of experimentation effects. In Panel A, we construct the simple difference in GDP per capita, among the experiment sites, before and after the policy trial; in Panel B, we further control for provincial pre-trends in GDP per capita; in Panel C, we estimate experimentation effects using a generalized synthetic control approach, where each experimentation site is matched with a weighted average of counties that shares a similar 3-year pre-trend, and minister and year fixed effects are included. Coefficients and robust standard errors are reported in the top-left corner of each panel.



Panel A Panel B

Figure 5: These plots demonstrate how policy effects shrink between the experimentation and roll-out stages. In Panel A, we plot policy effect during national roll-out (y-axis) against experimentation effect of the same policy (x-axis). In Panel B, we compute the the difference between policy effect during national roll-out and policy effect during experimentation, take its ratio over the experiment effect, and plot its distribution.

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	652	2.1	19.0	42.02	5.01	43.0
National	613	2.1	19.7	43.72	5.17	41.9
× Completed	509	2.1	18.8	50.88	5.48	39.4
× Ongoing	104	2.1	23.9	8.65	3.58	56.3
Subnational	39	2.0	8.7	15.38	2.22	60.0
× Completed	35	2.0	9.2	17.14	2.16	59.3
× Ongoing	5	2.0	11.4	0.00	2.72	50.0
Panel B: **By policy domain**						
Resource, energy & environment	80	2.2	11.5	38.75	3.94	57.1
Market supervision	79	1.9	10.9	44.30	5.91	33.9
Agriculture	60	2.1	39.4	33.33	3.91	56.6
Education	56	2.3	39.2	46.43	5.43	28.3
Finance	53	1.8	6.2	47.17	8.50	40.6
Tax & fiscal policy	41	2.2	10.2	53.66	5.38	38.2
Population & health	38	2.3	21.2	47.37	4.57	36.1
Commerce & trade	36	2.1	17.0	41.67	6.34	23.1
Industry & information technology	35	1.8	25.2	37.14	6.87	24.0
Domestic affairs	31	2.3	15.7	29.03	4.12	36.0
Development & reform	29	2.0	23.0	37.93	4.25	60.0
Labor	22	2.5	9.9	45.45	4.61	55.6
Transportation	20	2.0	9.2	55.00	3.09	58.8
Others	33	1.9	34.2	66.67	5.27	40.0
Panel C: **By administrative level**						
Province-level	199	1.7	4.9	36.18	1.44	72.4
City and county-level	414	2.3	26.8	47.34	6.57	31.5

Note: This table reports the summary statistics for our policy experimentation sample. In Panel A, we present information on all 652 experiments, and disaggregate them by national experiments (613) and subnationl ones (39). In Panels B to F, we only focus on those national experiments.

Table 2: 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***	-0.014***	-0.002*	-0.003
	(0.001)	(0.000)	(0.000)	(0.003)	(0.001)	(0.002)
# $\times$ career incentive				0.036***	0.009***	0.011***
				(0.006)	(0.003)	(0.003)
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
# of clusters	1973	1973	1973	1973	1973	1973
Mean of Dep. Var	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. The average number of policy experiments within each prefecture-year-domain grid is 0.218, and the standard deviation is 0.541. Career incentives are measured as the ex-ante probability of promotion projected by the start age of tenure and hierarchical level (mean=0.481, s.d.=0.075). 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 the county level.

Table 3: Naive evaluation of policy experimentation

	National roll-out		
	(1)	(2)	(3)
Panel A: Land revenue windfall			
Land revenue (instrumented)	0.008***	0.006***	0.009***
	(0.002)	(0.001)	(0.001)
First stage F stats	670.34	636.36	622.90
# of obs.	66,128	66,128	66,128
# of clusters	1644	1644	1644
Mean of DV	0.612	0.612	0.612
Experiment Year FE	Yes	Yes	Yes
County FE	No	Yes	Yes
Ministry FE	No	No	Yes
Panel B: Politicians' incentive changes due to political rotation			
Rotation	0.004	0.016	0.017
	(0.019)	(0.014)	(0.014)
Positive rotation $\times\Delta$ Incentive	0.679***	0.539***	0.508***
	(0.089)	(0.090)	(0.093)
Negative rotation $\times\Delta$ Incentive	-0.496***	-0.459***	-0.407***
	(0.162)	(0.132)	(0.141)
# of obs.	3899	3899	3899
# of clusters	27	27	27
Mean of DV	0.321	0.321	0.321
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 as a national policy. 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 (in logarithm) received by the local government. The average log land value is 5.27, with a standard deviation of 3.97. Standard errors are clustered at the county level. We report the first stage results in Appendix Table A.14. 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. An average positive rotation is accompanied with an incentive increase of 0.079 (s.d.=0.076). An average negative rotation is accompanied with an incentive drop of 0.055 (s.d.=0.061). The standard errors are clustered at the province level.

Table 4: Similarity with experimentation sites and effects of policy roll-out

	Growth of GDP per capita		
	(1)	(2)	(3)
Panel A: Selection of experimentation sites			
M-distance in socioeconomic conditions	-0.004***	-0.004***	-0.003***
	(0.001)	(0.001)	(0.001)
# of obs.	94,772	94,772	94,772
# of clusters	2064	2064	2064
Mean of DV	0.102	0.102	0.102
Panel B: Endogenous efforts during experimentation			
M-distance in politicians' career incentives	-0.001***	-0.001***	-0.0002
	(0.0002)	(0.0003)	(0.0002)
# of obs.	55,940	55,940	55,940
# of clusters	1464	1464	1464
Mean of DV	0.088	0.088	0.088
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 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 the 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.26, 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_281476529291118.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.27 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.26 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:

$$\hat{y}_{pt} = \Phi^{-1}\left\{\hat{\alpha} \cdot startage_{pt} + \hat{\beta} \cdot level_{pt} + \hat{\gamma} \cdot startage_{pt} \times level_{pt}\right\}. \tag{4}$$

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.29 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 1999–2006, 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 three different sources: the China Strikes project (2002-2011), the China Labor Bulletin (2012-2020), and the Global Database of Events, Language, and Tone (GDELT, 2014-2020), some of the largest databases on political events. See https://chinastrikes.crowdmap.com/ for the China Strikes website, https://clb.org.hk/en for details of the China Labor Bulletin data, and 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:

$$y_{mt} = \sum_k D^k_{mt} \cdot \beta_k + \delta_m + \theta_t + \varepsilon_{mt}, \tag{5}$$

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.28 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 Potential reasons for positive selection

D.1 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 representative 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).

D.2 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 Re-

form 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 **??**). 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) 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:

$$y_{pt} = \alpha \cdot Incentive_{pt} + X'_{pt} \cdot \beta + \delta_p + \theta_t + \varepsilon_{pt}, \tag{6}$$

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 Appendix Table A.27, Panel A, when the prefectural leaders have stronger promotion incentives, the corresponding localities engage in significantly more policy experiments. This result is robust if we adopt an alternative definition of career incentives exploiting the jump of promotion probability at the age cutoff 58. In Appendix Figure A.29, we estimate a standard regression discontinuity model,

$$Y_{it} = \alpha + \tau D_{it} + \beta_1(X_{it} - 58) + \beta_2 D_{it}(X_{it} - 58) + \varepsilon_{it}$$

where Y_{it} is the total number of policy experiments that prefecture i carries out in year t, and D_{it} is a dummy variable capturing whether the prefecture leader hits the age limit.

Figure A.29 plots the results. Consistent to our findings in Table A.28, we find the

politicians with lower promotion incentives participates in less policy experiments. The RD point estimate is -0.267** (0.122).

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.28). Moreover, such effects of promotion incentives are almost entirely driven by policy experiments initiated by M-form ministries (see in Appendix Table A.27). 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 experimentation. 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:

$$y_{mpt} = \alpha \cdot Connection_{mpt} + \delta_{mp} + \theta_t + \varepsilon_{mpt}, \tag{7}$$

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; 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 A.23, Panel A, when a region becomes connected to a minister, the number of experiments assigned to that region increases immediately by 28.8%.[2] 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

2. In Appendix Figure A.25, 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.

are selected via voluntary participation (see Appendix Table A.23). This suggests that the political patronage in experimentation site selection works through top-down favoritism.

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 F.

E 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).

E.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 **??**. 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.[3] 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.

3. 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.[4]

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.30. 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.31. 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.[5]

The results from these simulations are displayed in Figure A.32. 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.

4. 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.

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

E.2 Simulations with welfare considerations following Narita 2021

Overview Second, we examine how the optimal experimentation design would change if the decision maker incorporates considerations over the welfare of the experimentation sites. Narita 2021 demonstrates that deviation from full randomization may be justified when the sample size is finite, and there exists sufficiently large heterogeneous treatment effects as well as heterogeneous welfare from receiving the treatment policy (either captured as experimentation subjects' willingness to pay, or benevolent social planners' welfare weights across subjects). We again simulate the optimal experimentation design, parameterizing the model based on the experimentation setup and estimated heterogeneous treatment effects from Section **??**. We find that the central government would have to place almost the entirety of its welfare weights on the locations that were selected as the experimentation sites in the early waves in order to justify the observed degree of positive selection, suggesting that the observed positive selection of experimentation sites could be optimal only if extreme *ex ante* inequality is inherent to central government's objective function.

Narita 2021 presents a model that incorporates the welfare of subjects into the experimental design process. Specifically, alongside the DM's priors on predicted treatment effects, subjects also have their own willingness to pay for each treatment. This information is used to construct the Experiment-as-Market (EXAM) design, which provides a price-discriminated competitive equilibrium such that:

1. Subjects are given a budget b, so that the price of a treatment $\pi_{te} = \alpha e + \beta_t$ is decreasing in predicted treatment effect e for each treatment t (which in our case is simply a treatment and control). Thus, a solution must have $\alpha < 0$

2. Subjects maximize utility, satisfying

$$(p^*_{it})_t \in \arg\max_{p_i \in P} \Sigma_t p_{it} w_{it} \quad \text{s.t.} \quad \Sigma_t p_{it} \pi_{te_{ti}} \leq b$$

   where i indexes a subject, w_{it} is the subject's willingness to pay for a treatment, and p_{it} is the probability that the subject i receives treatment t.

EXAM holds two nice properties, namely that (1) no other experimental design Pareto dominates EXAM in expected treatment effect or WTP, and (2) any parameter estimable without bias under RCT, including the ATE, is also estimable without bias under EXAM. The general algorithm to find the EXAM equilibrium is laid out in Appendix 3B of Narita 2021.

We follow the same algorithm to find the equilibrium probabilities of treatment, using treatment effects as described in step 1 of Section E.1 and willingness to pay following:

$$w_{c,t} = \begin{cases} 0 & \text{for } t = \text{control} \\ \beta_c + \eta_c \qquad \eta \sim U[-\beta_{max}, \beta_{max}] & \text{for } t = \text{treatment} \end{cases}$$

Furthermore, counties are endowed with a budget b_i whose valuation follows the pdf of the distribution $\text{Beta}(\delta, 10-\delta)$ where $1 \leq \delta \leq 10$. Given that the beta distribution only has a support on $[0,1]$, GDP values are scaled according to the formula: $GDP_{scaled} = \frac{GDP - GDP_{min}}{GDP_{max} - GDP_{min}}$ and the budget allocated to a county i is therefore $b_i = \text{Beta}(\delta, 1-\delta)(GDP_{scaled,i})$.

Higher values of δ give a greater budget (and place more welfare weight) on counties with a greater GDP per capita, while a $\delta = 5$ valuation has no bias towards wealthier or poorer counties.

We generate 1,000 sets of WTPs using this data generation process and compute the set of optimal $(p^*_{c,t})$ that clear the market. For each set of optimal prices, we generate 1,000 experimental assignments based on the implied probability of treatment and compute the mean t-statistic for this set of WTPs. Appendix Figure A.33 shows optimal t-statistics for simulations calibrated using three different policy experiments conducted in China.

F 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 D.2). For the baseline, we estimate the following econometric model using policy-prefecture level data:

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

Appendix Table A.30 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.34, Panel B shows the distribution of t statistics of the representative test, as described in Section 5.1, 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.34, 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.

Moreover, 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.34, Panel C. This specification is most similar, in general ideas, to the regression presented in Table A.30, confirming the idea that all distortion factors we identified explain almost half of the selection bias of policy experimentation.

G Political incentives and differentiation during experimentation

In addition to the increased domain-specific fiscal expenditure during experimentation (as presented in Section 6), we examine whether local politicians with stronger career incentives differentiate their implementation activities more during trial policy implementation. Differentiation can signal effort and potentially earn political credit as a "model experimentation site."[6]

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 SimilarityNet, Baidu's state-of-the-art algorithm for short-text similarity scores.[7] This exercise follows Bertrand et al. (2020) and Acemoglu, Yang, and Zhou (2021) in spirit.

Short Text Semantic Matching (SimilarityNet, SimNet) is a framework for calculating the similarity of short texts. With a standard input-representation-matching layer structure, it can calculate the similarity score based on two texts input by the user. It mainly includes BOW, CNN, RNN, MMDNN and other forms of core network structure, providing semantic similarity computation training and prediction framework. It is widely used in actual scenarios including information retrieval, news recommendation, intelligent customer service, and so on.

To motivate our measure, we take the Carbon Emission Trading (CET) policy (See Section A.2.1 for details) as an example. In this example, we start with the *Interim Measures for the Administration of Carbon Emission Trading in Shenzhen (2014)*, one of the first batch of regulations issued by the local government after the CET experimentation. This stands as a good example because it enables us to compare vertically with a later version of the exact policy issued by the same locality: *Measures for the Administration of Carbon Emission Trading in Shenzhen (2022)* (similarity score = 0.91), and horizontally with the (first) CET regulation documents issued by other localities that participates in the experimentation. (e.g. Fujian, 2018, similarity score=0.869; Chongqing, 2023, similarity score=0.885). This echoes with our prior that across-regional variation of policy details is usually larger than within-region across-time variation, because the latter comes with some bureaucratic persistence.

As a placebo exercise, we also test the pairwise similarity between our baseline policy document with a set of "irrelevant" policies. Albeit discussing a variety of themes, it could be the writing of the bureaucrats and secretaries, or the irrelevant higher-level slogans that drives the similarity pattern. To rule out this possibility, we take a random sample of 100 irrelevant policy documents and plot the distribution of pairwise similarity indices. Figure A.35 shows that non of the placebo tests generate a similarity index larger than the

6. When a policy experiment turns into a national policy, the central government typically selects one of the better-performing experimentation sites as a model site, whose experience in implementing that policy will be promoted to the rest of the country.

7. Versions: `paddlepaddle==2.3.2`, `simnet_bow==1.2.1`

baseline document pair (indicated by the dashed grey vertical line). It suggests that our measure does seem to be capturing the most important margin relevant to our research.

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 politician's career incentives, as in Section D.2; 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 locality 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 Appendix D.2, 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 Appendix Table A.13. We observe that, when local politicians have strong career incentives, they tend to differentiate more than their colleagues in terms of implementation details, reflecting an increase in local politicians' efforts to achieve good results in the experiment.[8]

8. Such differentiation may be sub-optimal — for example, if policy solutions that are proven effective had already been tried out by their peers in previous waves of experimentation. However, we do not have evidence to evaluate the optimal level of differentiation during policy experimentation.

H Accounting for the magnitude of positive selection selection and strategic efforts' impact on national policy roll-out due to naive inference

We have shown that the central government is not fully sophisticated in evaluating policy experiments: it cannot fully separate the influences of exogenous shocks to the experimentation sites that are unrelated to the policy experiments, such as windfalls of fiscal revenue and unexpected boosts of local political incentives. In this section, we ask, if the central government exhibits the same levels of errors with respect to the previously documented patterns of positive site selection (Section 5) and endogenous effort (Section 6), to what extent would these patterns affect national policy roll-out decisions?

Assuming that the central government exhibits the same levels of mis-attribution errors for experimentation site selection and endogenous efforts, as they do for exogenous shocks such as fiscal windfalls and political rotations, and assuming that all the estimated effects can be extrapolated linearly, we conduct simple back of the envelope calculations to gauge the magnitude of potential biases in policy learning and policy choices originating from biased non-representative policy experimentation.

We start by estimating the elasticity of policy roll-out with respect to fiscal expenditure, leveraging our estimates from Table 3, Panel A, Column 3. During our sample period, land revenue accounts for 22.7% of total fiscal revenue. Our IV estimate thus informs us that a 1% increase in land revenue in an experimentation county, which translates into a 0.227% increase in total fiscal revenue, leads to a 0.009 percentage point increase in roll-out probability. We thus know that a 1% increase in fiscal revenue in an experimentation county is correlated with a 0.04 percentage point increase in roll-out probability. We then multiply this number by the median number of counties participating in a policy experiment (18), thus obtaining an estimated elasticity of policy roll-out with respect to fiscal revenue increase of 0.72.

Using this elasticity, we calculate the impacts of fiscal revenue increase from two sources - positive selection that occurs prior to the start of the policy experiment, and endogenous effort during the experimentation period. In the former scenario, we focus on the county-level experiments to consistently use the estimated coefficient from the land revenue exercise. There's a 20.5% difference between the average fiscal revenue in the pre-experimentation period between the treated units, versus the control units. Linking this number to the estimated elasticity, the inflation in national roll-out rate related to the observed level of positive selection, through the channel of fiscal revenues, would be $20.5 \times 0.72 = 14.76$ percentage points.

For the second channel (endogenous fiscal input), we use the estimates in Appendix Table A.31 to compute the average increase in total fiscal expenditure driven by experimentation participation. We observe a 8.1% increase in total fiscal expenditure for each additional policy experiment. Linking this number to the estimated elasticity, we calculate that, the total bias related to the observed level of endogenous efforts, through the channel of fiscal revenues, would be $8.1 \times 0.72 = 5.83$ percentage points.

Similarly, we also calculate the impacts of political incentives on policy roll-out. Ac-

cording to the interaction-term coefficient in Table 3, Panel B, Column 3, a 0.01 unit increase in an experimentation site's local political incentives would increase the overall roll-out likelihood by 0.0015. Linking this number to the average number of experimentation sites per experiment (21), as well as the baseline average difference in political incentives between experimentation and non-experimentation sites (0.006), we calculate that the positive selection in political incentives would inflate policy roll-out rate by $15.1 \times 0.006 \times 21 = 1.9$ percentage points.

A cautionary note is that, in addition to the typical linearity assumptions made in such back-of-the-envelope calculations, the exercises above also hinge on the assumption that the central government exhibits the same levels of errors for experimentation site selection and endogenous efforts as they do for exogenous shocks such as fiscal windfalls and political rotations.

I Additional figures and tables

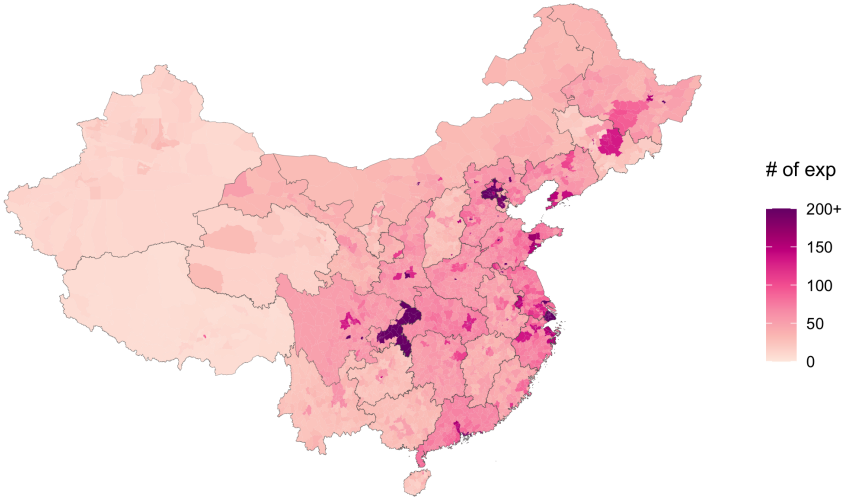


Figure A.1: This county-level map plots the spatial distribution of policy experiments in China. We add a count towards a county if either itself or its corresponding prefecture/province serves as an experimentation site for each policy experiment.

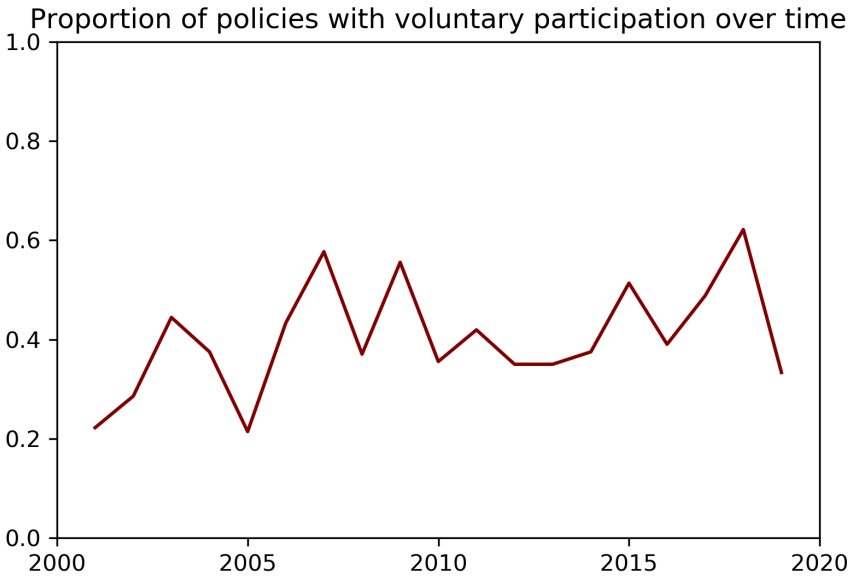


Figure A.2: This figure plots the share of policy experiments in each year that has a voluntary sign-up process for experimentation sites. We look for keywords and signs of voluntary sign-up in the first / main central document of each policy experiment.

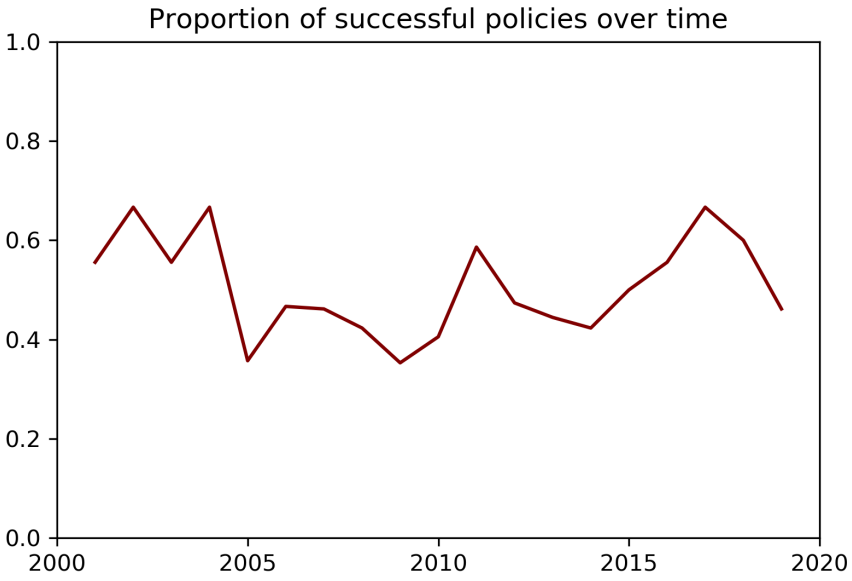


Figure A.3: This figure plots the share of successful policy experiments in each year. A policy experiment is defined as a "success" if we see evidence from a central government document that it eventually rolled out to the entire nation.

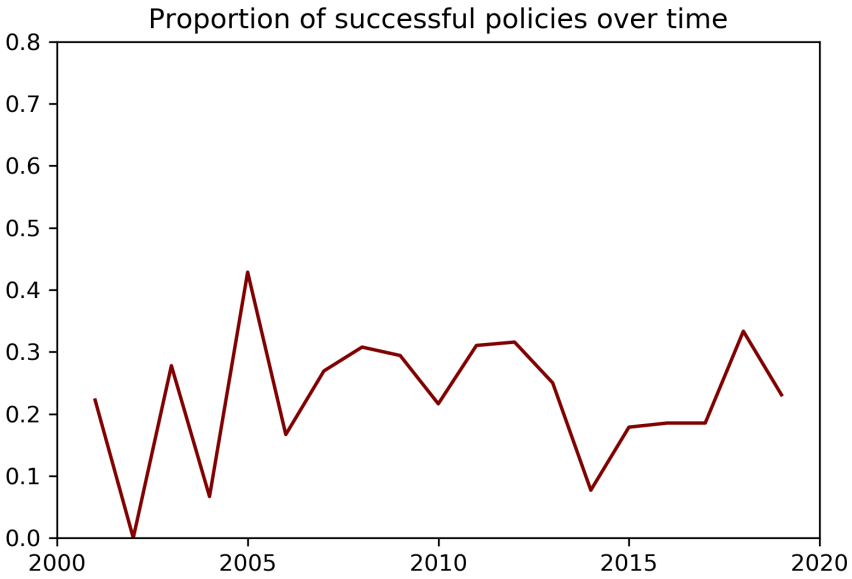
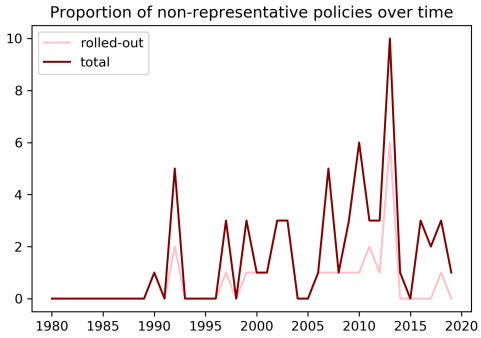
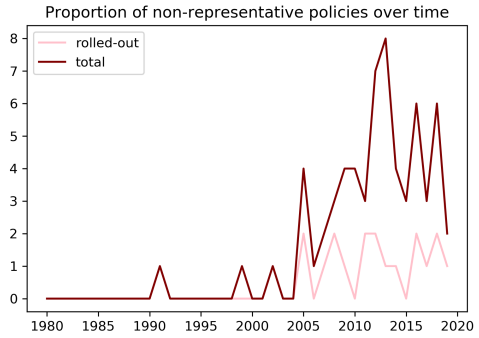


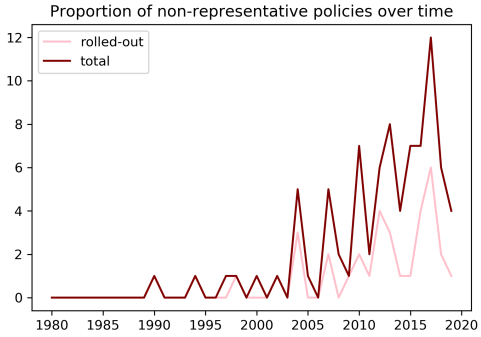
Figure A.4: This figure plots the share of successful policy experiments in each year. A policy is defined "success" if it is adopted by 2/3 of the provinces during and after the experimentation.



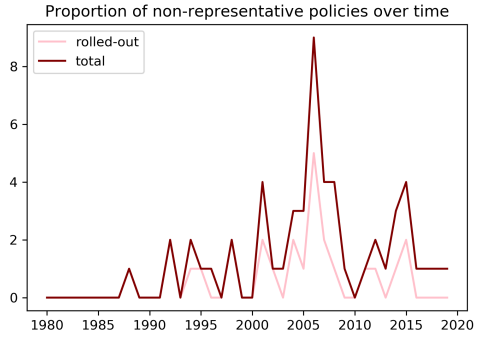
Education



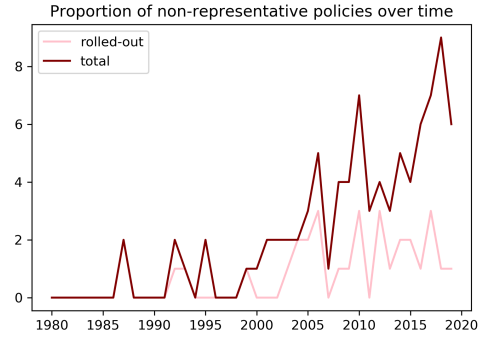
Agriculture



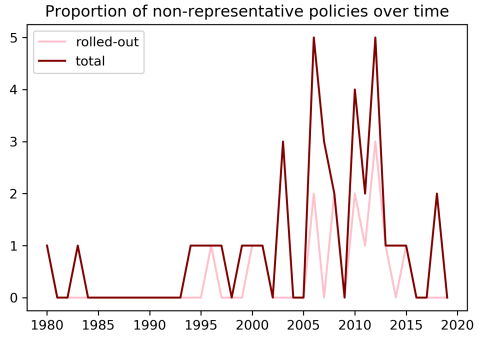
Market supervision



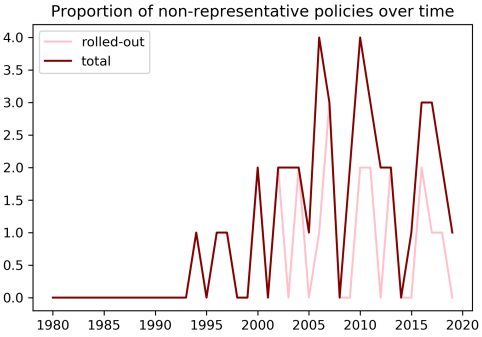
Finance & economics



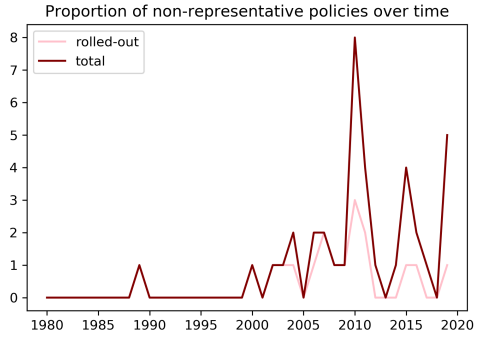
Natural resources & environment



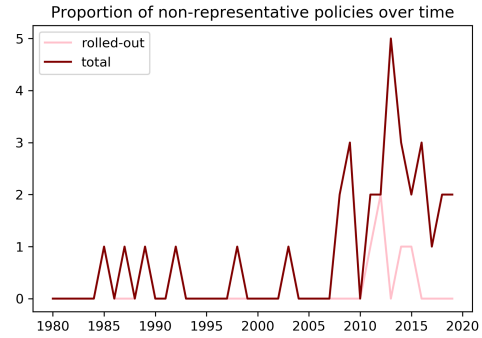
Business & commerce



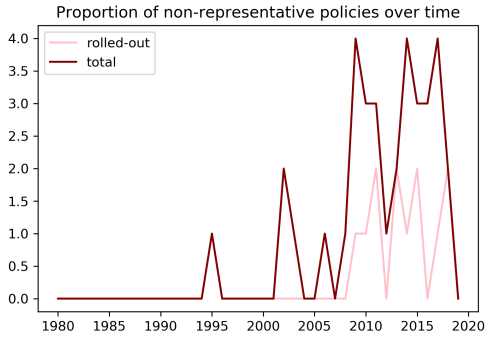
Government finance & taxation



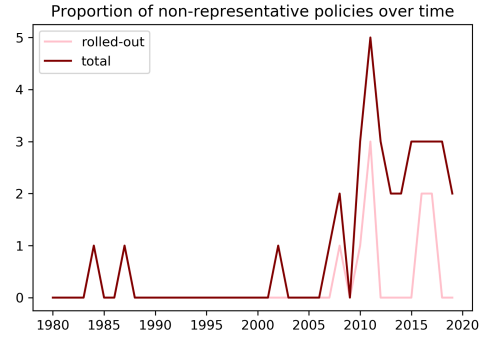
Population & health



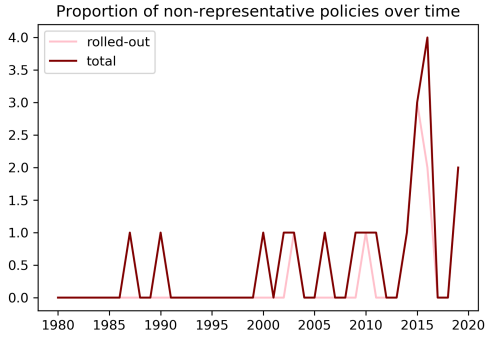
Domestic affairs



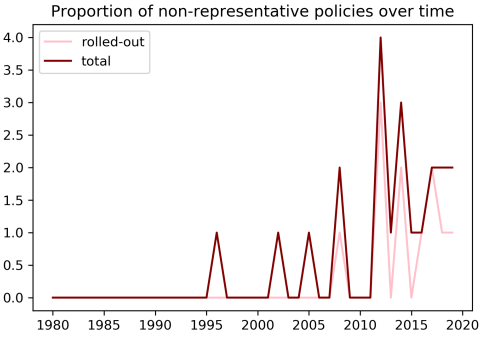
Industrial information



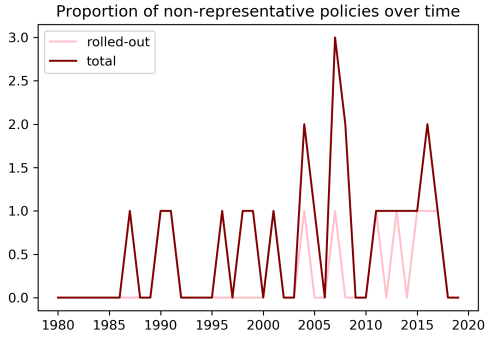
Development & reform



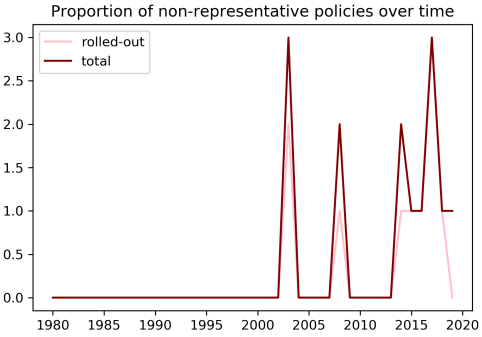
General purpose



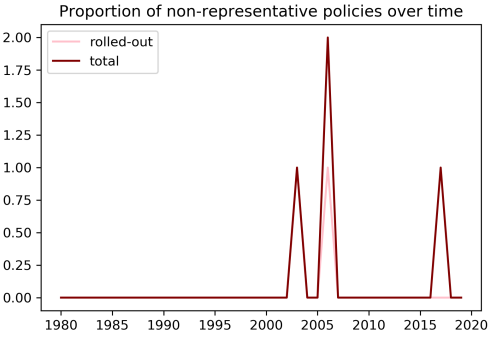
Transportation



Labor & personnel



Judiciary & supervision



Media

Figure A.5: These figures plot the count of policy experiments over time, by policy domains.

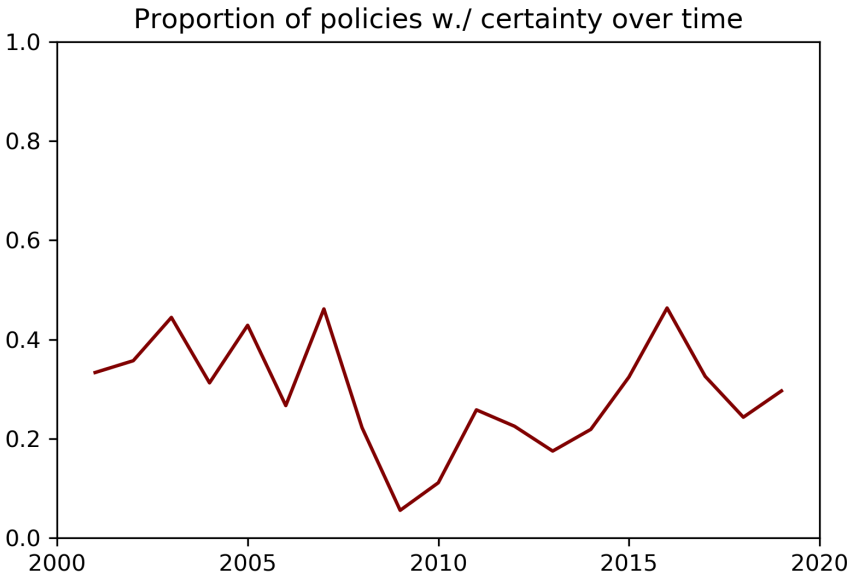


Figure A.6: This figure plots the share of policy experiments in each year that has detailed timelines of roll-out delineated in the first and main experimentation document. We consider these experiments as relatively more *ex ante* certain.

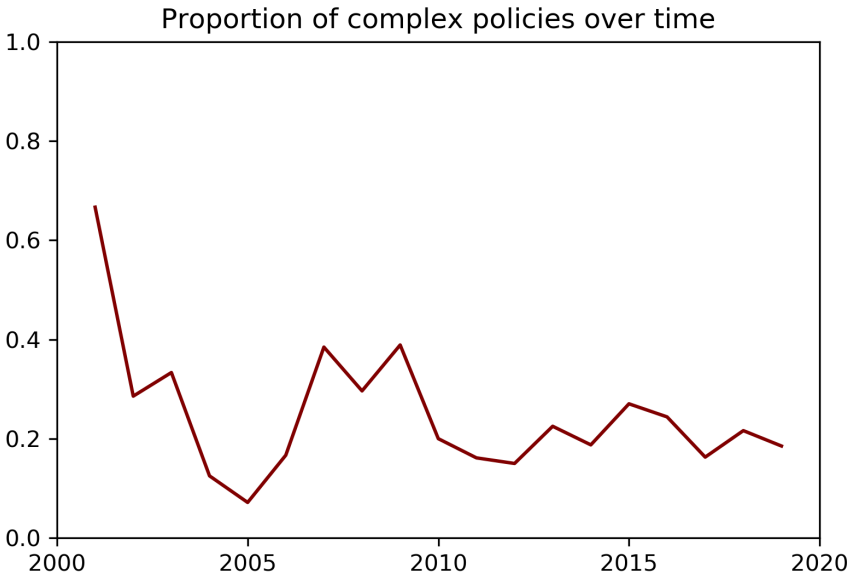


Figure A.7: This figure plots the share of policy experiments in each year that requires multi-department cooperation. For those policies, more than 1 ministry posted policy documents to lay out details about the experiments.

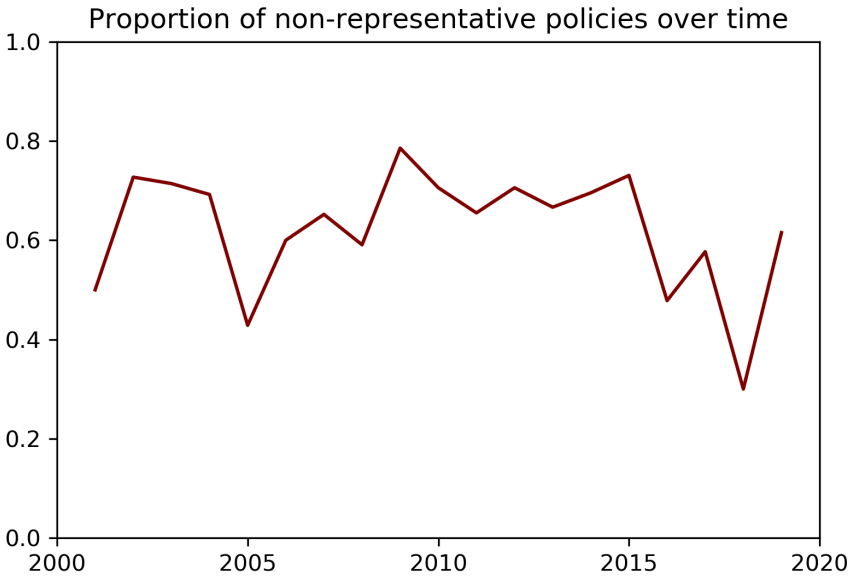
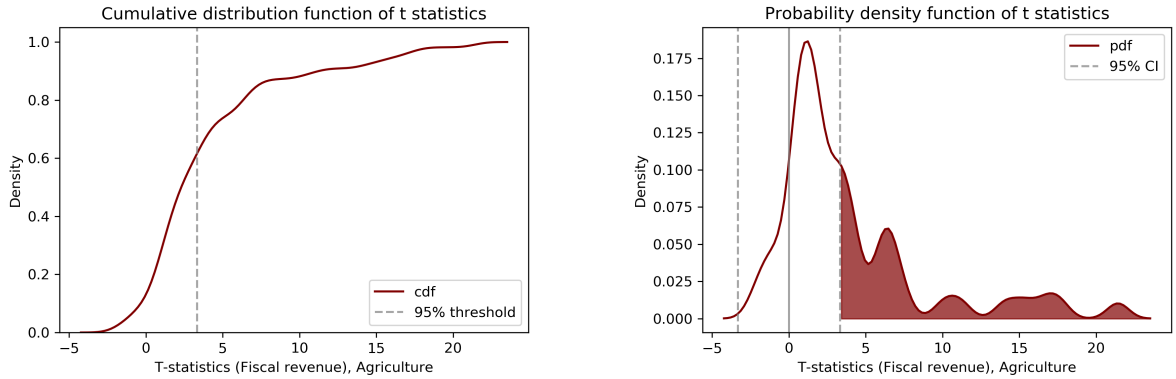
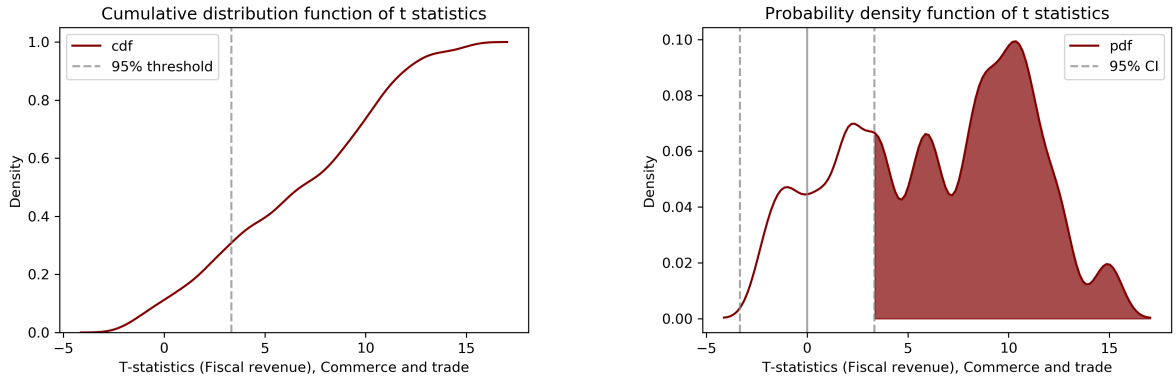


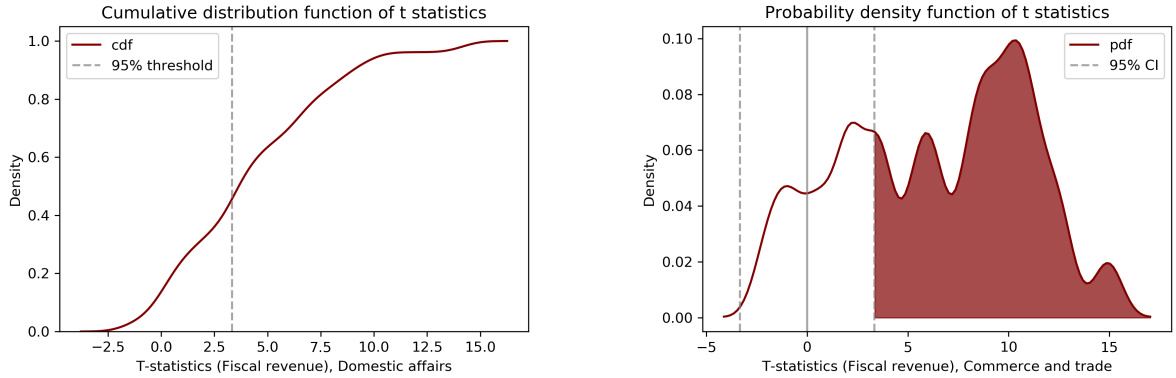
Figure A.8: This figure plots the share of non-representative policy experiments in each year. Non-representativeness is defined in Section 5.1.



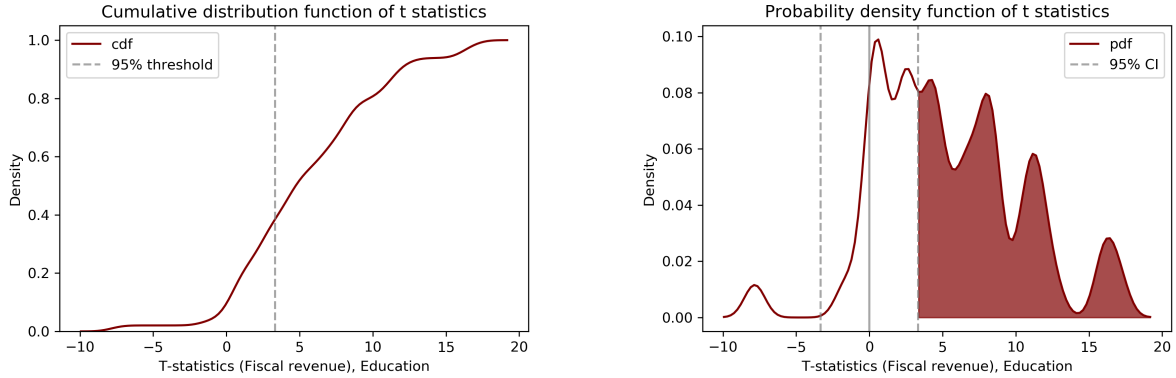
Panel A: Agricultural policies



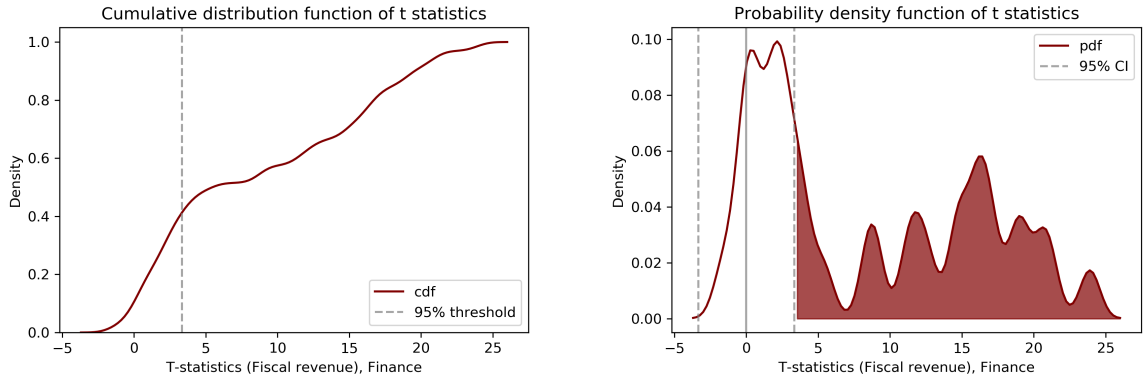
Panel B: Commerce and trade policies



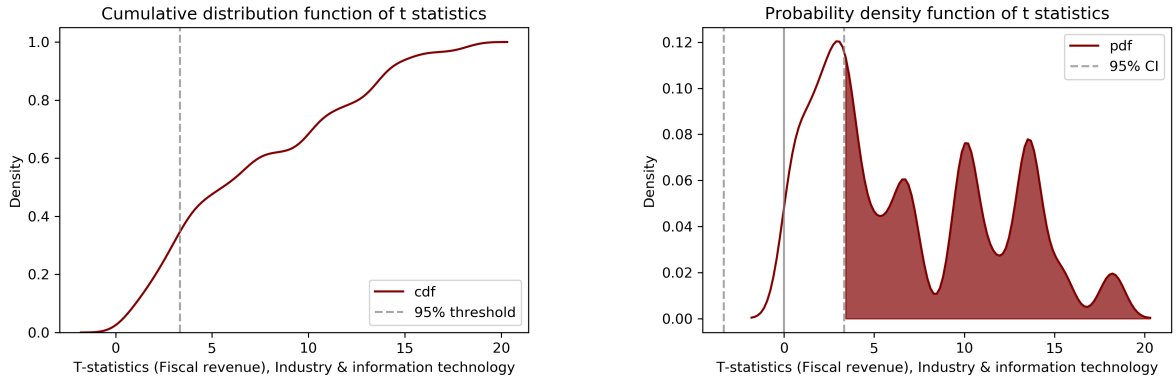
Panel C: Domestic-affair policies



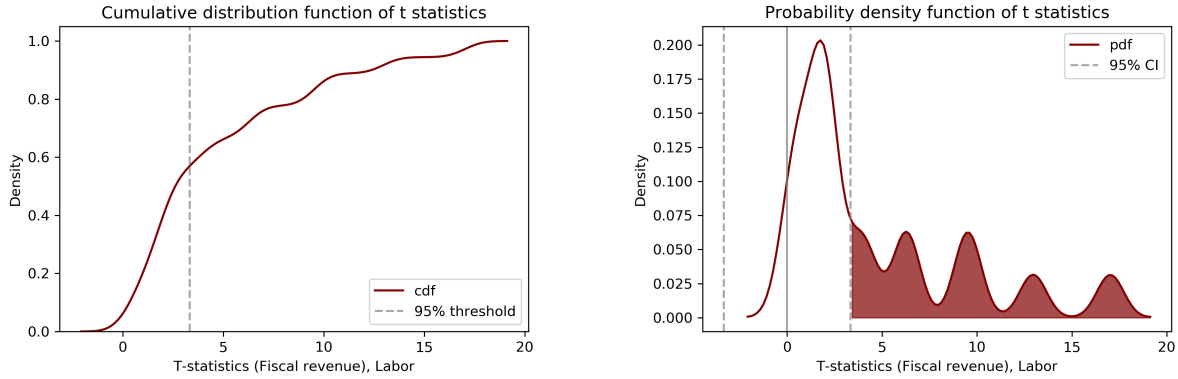
Panel D: Education policies



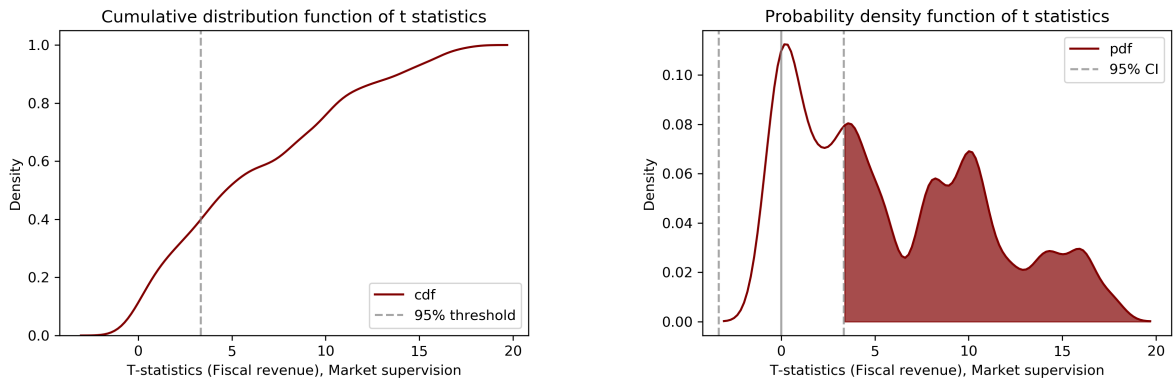
Panel E: Finance policies



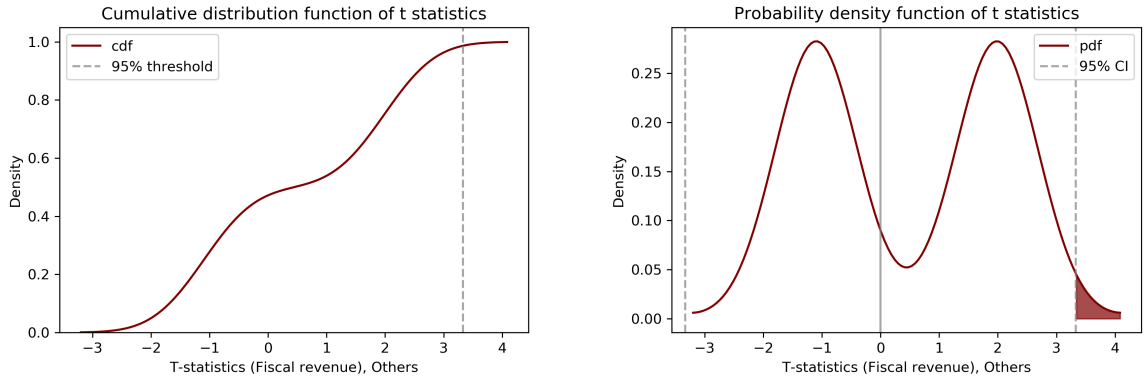
Panel F: Industry & information technology policies



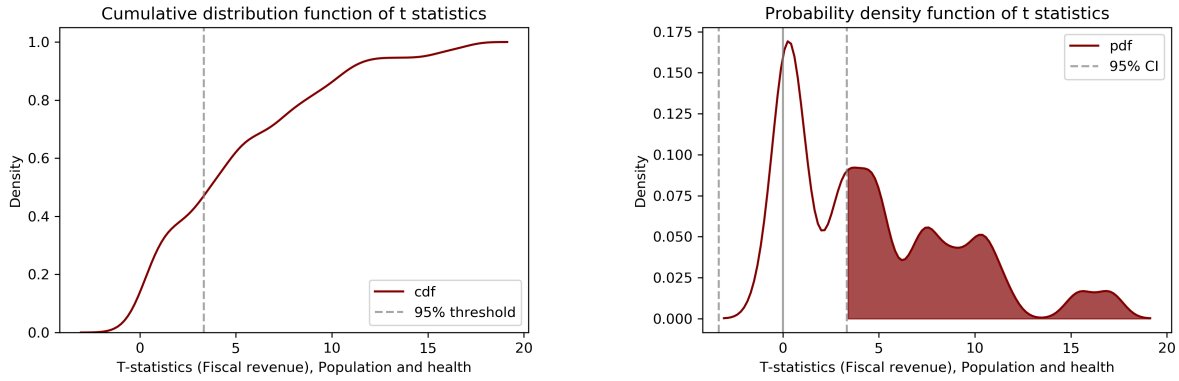
Panel G: Labor policies



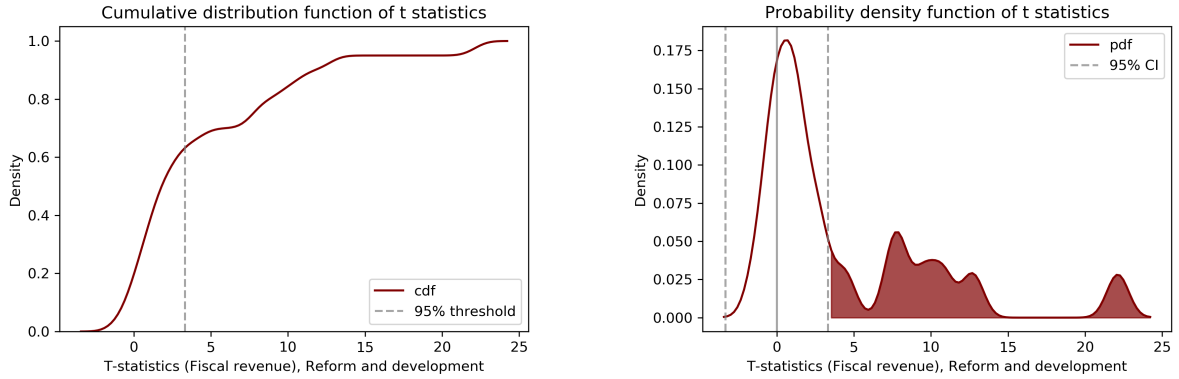
Panel H: Market supervision policies



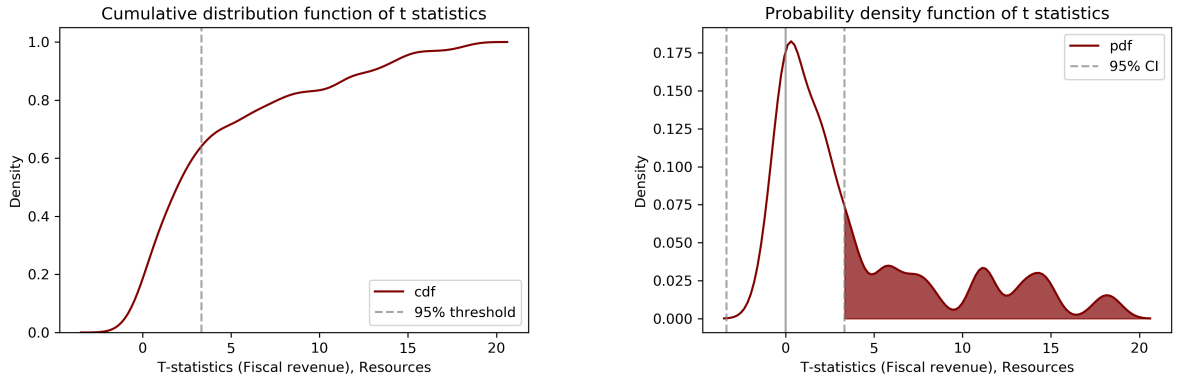
Panel I: Other policies



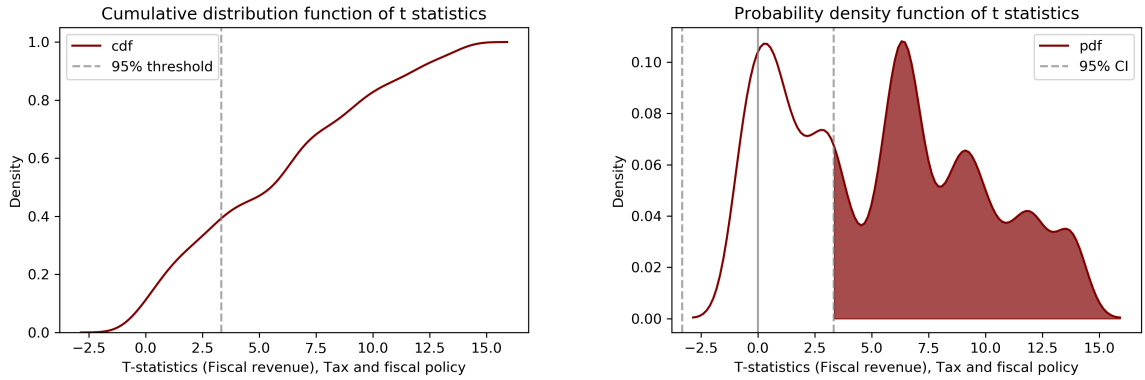
Panel J: Population & health policies



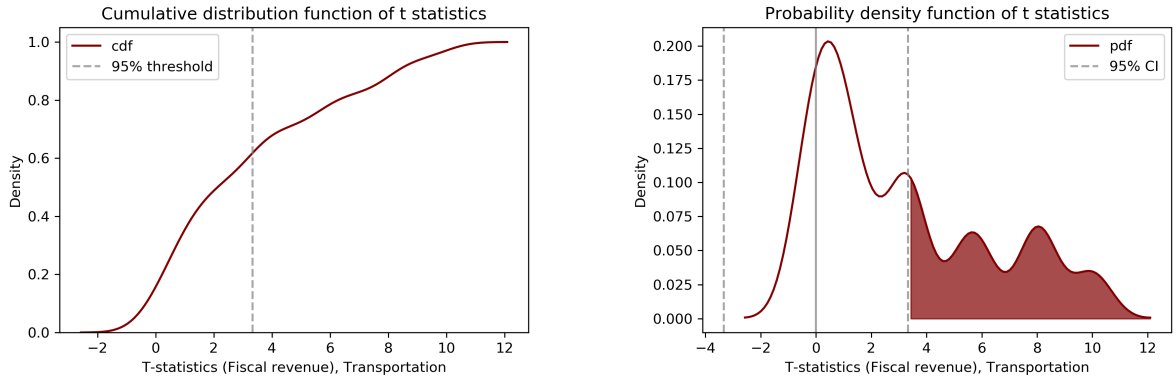
Panel K: Reform and development policies



Panel L: Resource-related policies

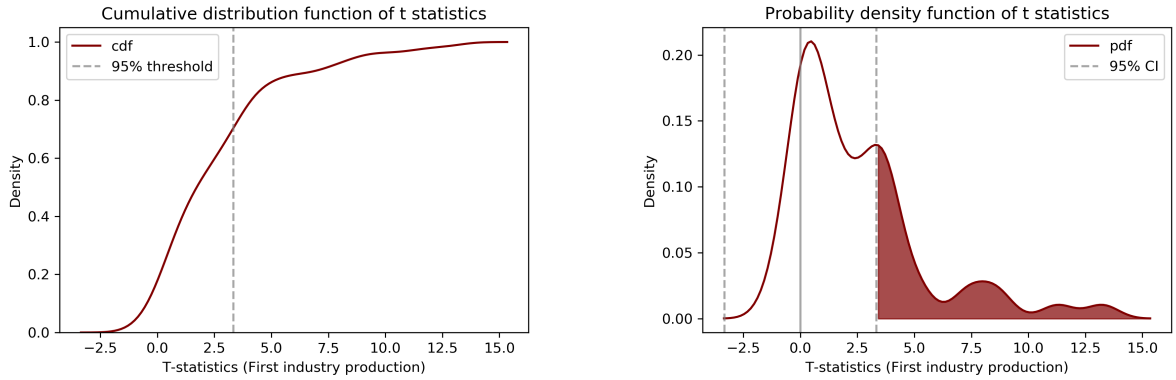


Panel M: Tax and fiscal policies

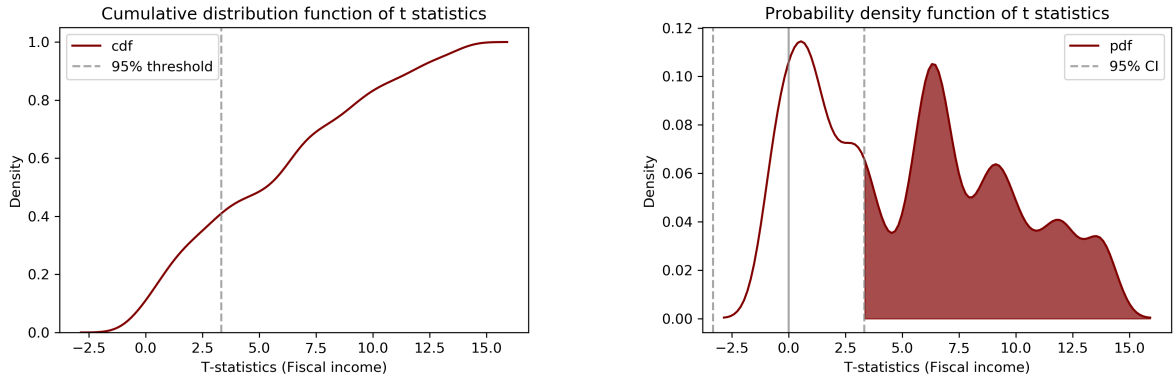


Panel N: Transportation policies

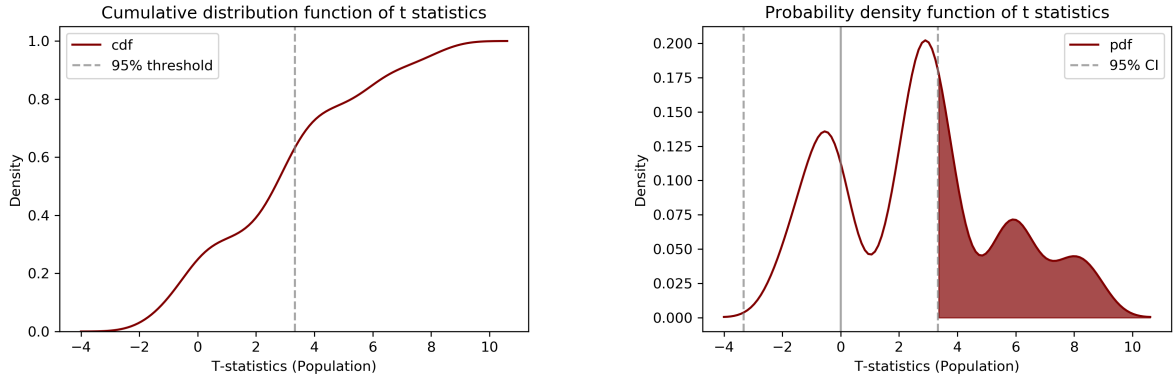
Figure A.9: Fiscal revenue t-tests on subsets of each of the 14 policy domains. We categorize each policy experiment to a main policy domain by looking for the ministry that published most relevant policy documents.



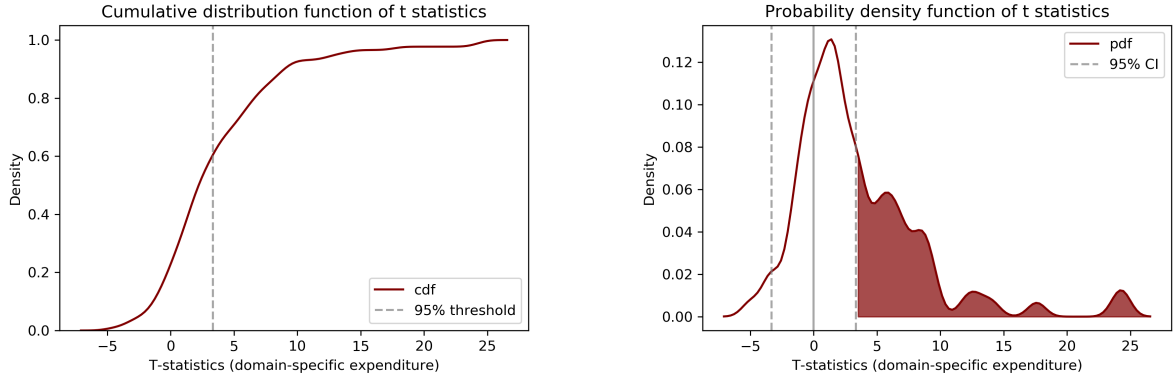
Panel A: Agricultural policies: test on agricultural contribution of GDP



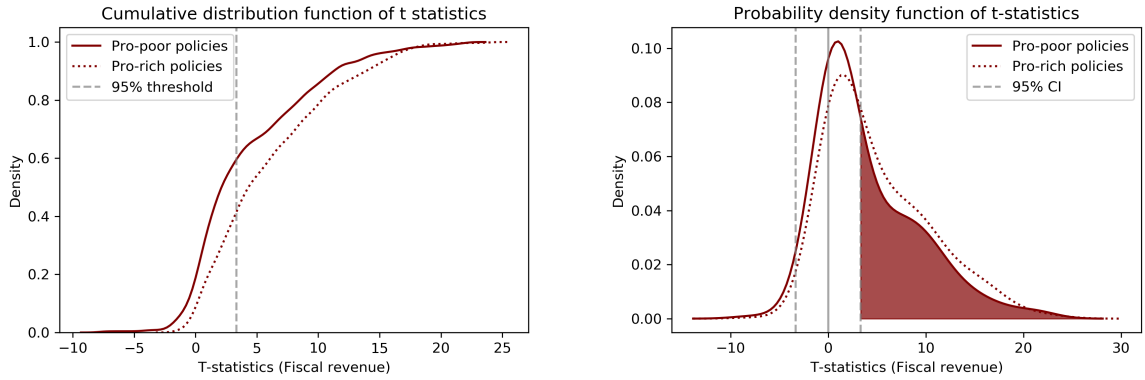
Panel B: Government finance and tax policies: test on fiscal income



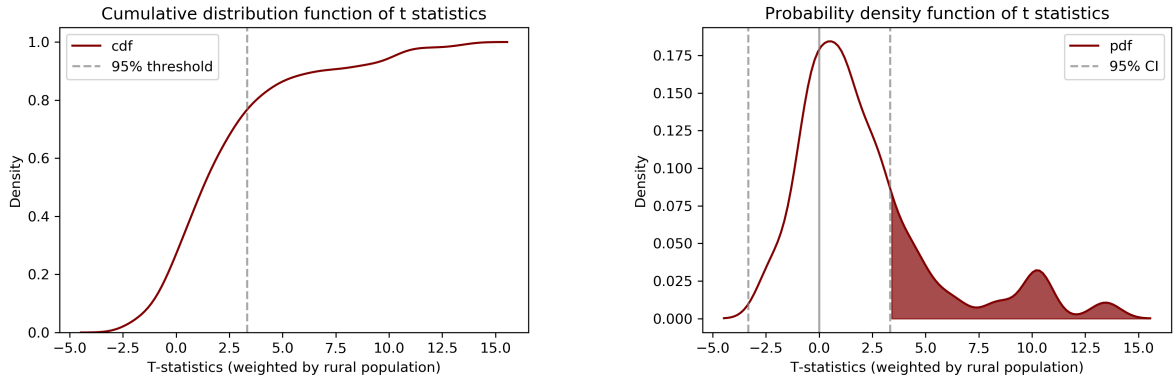
Panel C: Population and health policies: test on population levels



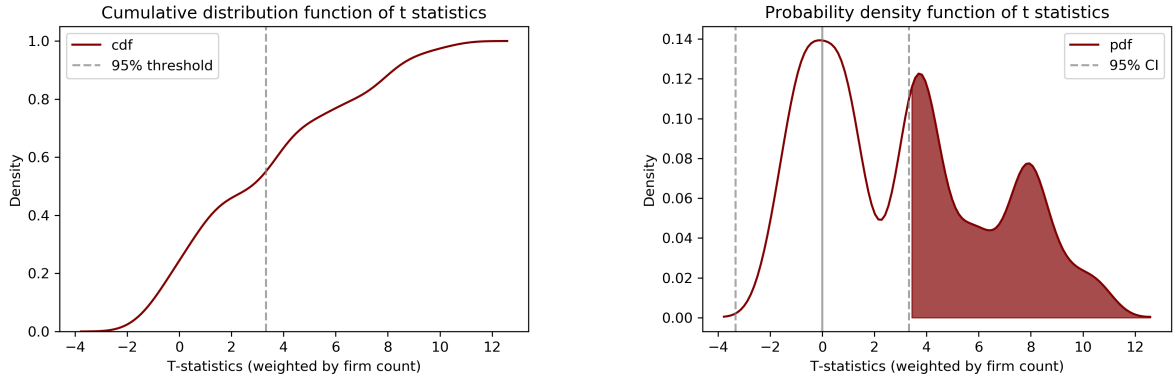
Panel D: Test with domain-specific fiscal expenditure



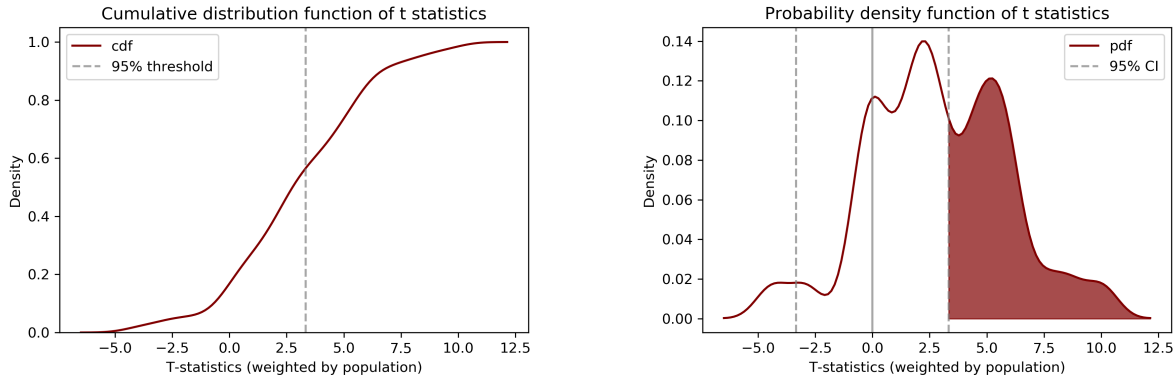
Panel E: pro-poor policy vs. pro-rich policy



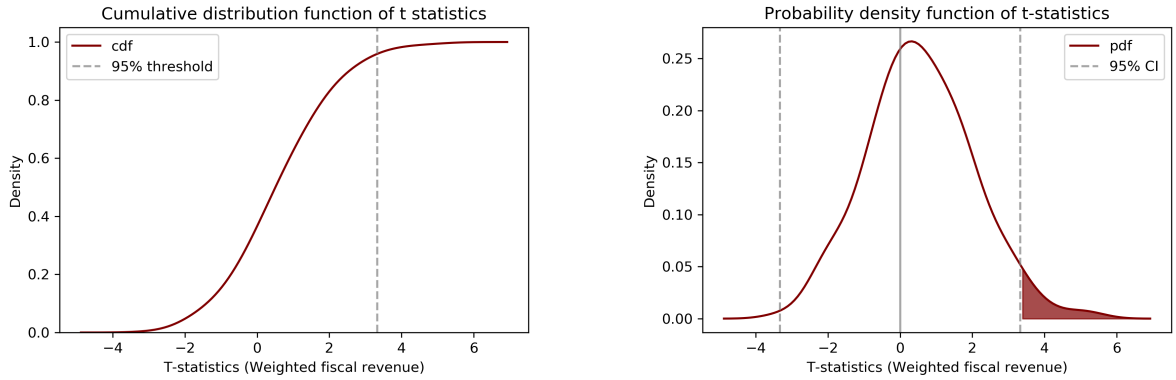
Panel F: Agricultural policies weighted by rural population



Panel G: Business policied weighted by the number of firms



Panel H: Healthcare policies weighted by population



Panel I: Population policies weighted by inward-migration rate

Figure A.10: Across panels we present a variety of robustness checks for positive selection patterns. The first three panels test on domain-specific outcomes on a subset of relevant policies. Specifically, in panel A we use agricultural outputs as outcome variable and focus on agricultural policies. In panel B we treat fiscal revenue as outcome variable and focus on tax policies. In panel C we test on population balance focusing on the subset of population and healthcare policies. In panel D, we return to the full sample. But instead of using locality-level fiscal revenue we focus on locality-domain-specific fiscal expenditure, which is more granular and relevant for the experimentation. In panel E, we distinguish between pro-poor and pro-rich policies and reproduce the baseline t-test. Pro-poor policies are defined as a subset of policies in which certain keywords such as "Anti-poverty", "Poor", "Rural areas", "Agriculture" are mentioned. The next four panels illustrates the robustness of our metric via the distribution of weighted t-stats. Specifically, in panel F we weight the fiscal revenue by the rural population of each locality and focus on the subset of agricultural policies. In panel G we weight the fiscal revenue by the number of firms focusing on the subset of commerce and business policies. In panel H we weight the fiscal revenue by population focusing on the subset of population and healthcare policies. And finally, in panel I we weight the full-sample fiscal revenue on inward-migration rate to account for the possiblity that differential weights are given to localities with more inward-migration trends. The level of inward migration is computed using 2000 and 2010 census. We only included 15-64 year olds in the sample and excluded students to compute prefecture-pair level migration flows, and divide them by the levels of population in 2010. Prefectural codes are mapped to 1991 versions.

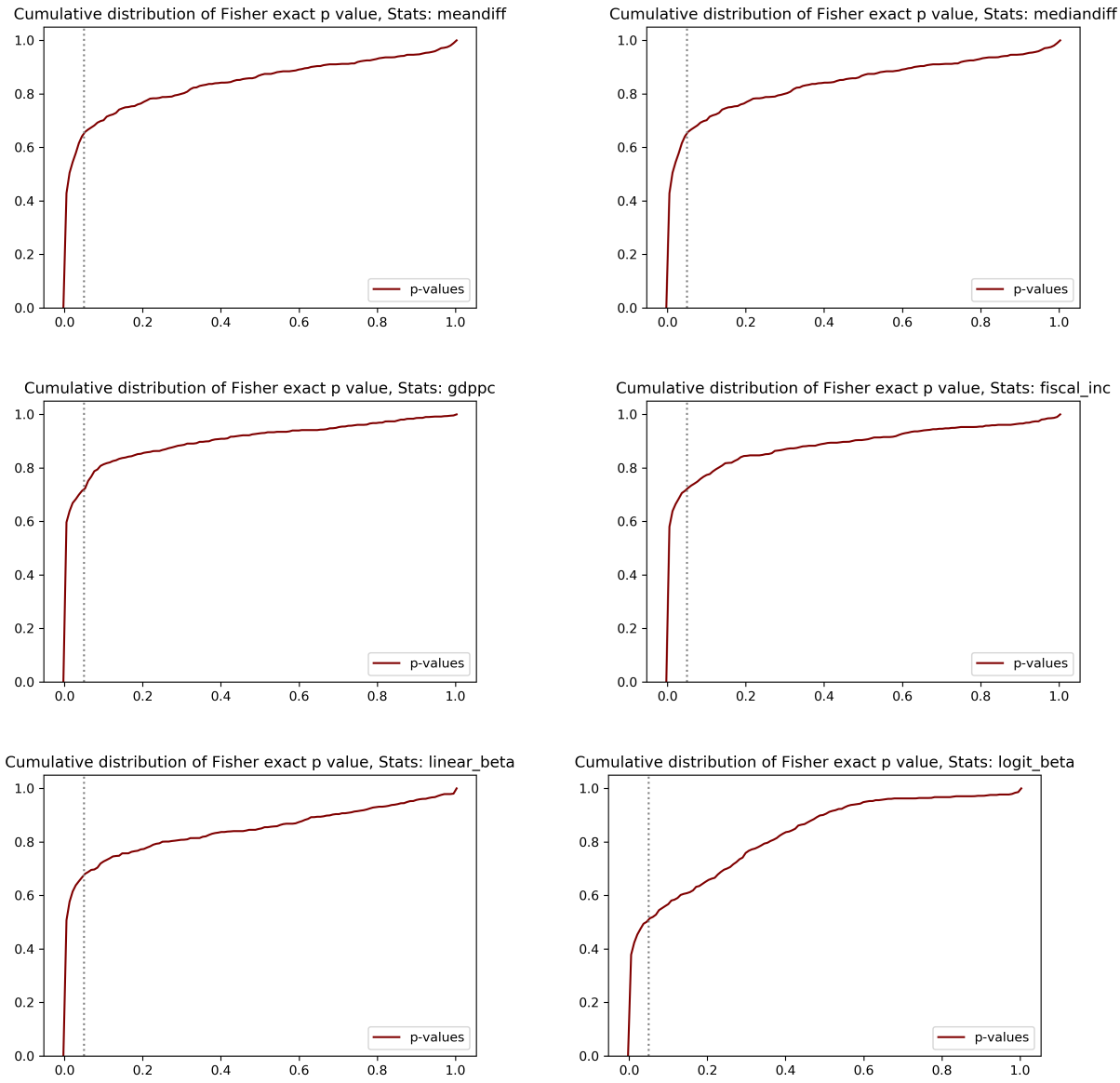


Figure A.11: These figures plot cumulative distribution of p-values from permutation tests. For each experiment, we randomly permute the treatment vector 500 times and compute the likelihoods of our observed statistic Θ exceeding its counterparts. In the upper panel, we use difference in means and difference in medians. In the middle panel, we use t-statistics on GDP per capita and fiscal revenue. In the last two figures, we choose Θ to be the regression coefficient $\hat{\beta}$ from the following specifications, respectively: $Z_i = \alpha + \beta X_i + \varepsilon_i$, and $Pr(Z_i) = \frac{1}{1+e^{\alpha+\beta X_i+\varepsilon}}$. For each perturbation of our treatment assignment mechanism, we can estimate a different $\tilde{\hat{\beta}}$. If pre-experimentation characteristics have enough predictive power on site selection, $\hat{\beta}$ estimated by observed treatment vector Z should be greater than most of the $\tilde{\hat{\beta}}$ estimated by perturbed treatment vector $\tilde{Z}$. The corresponding statistic should follow a uniform distribution across treatment assignment mechanisms, so the p-value equals the percentage of extreme values throughout permutation that are greater than the original $\hat{\beta}$. The dotted lines indicate cases where p-value $= 0.05$. We can reject at least 60% of all experiments being randomly selected.

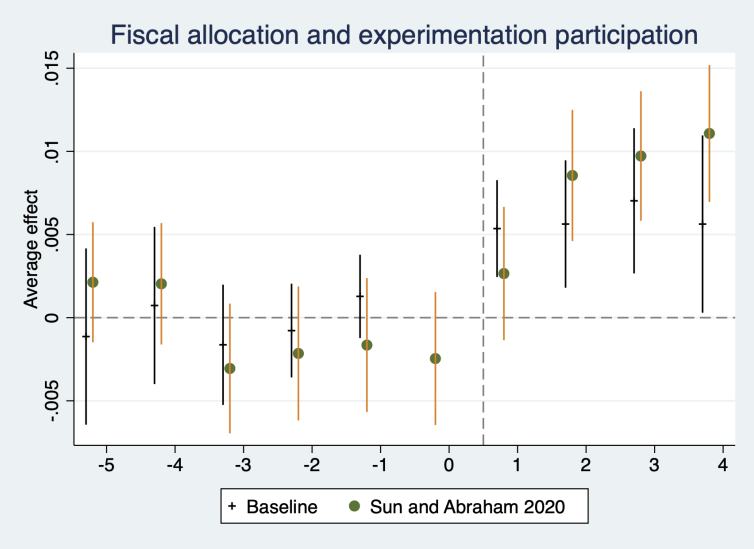


Figure A.12: This figure plots the event study estimates on the increase of fiscal expenditure in a specific domain after the county participated in an experiment in the corresponding domain. The sample period is 1993-2006. If there are multiple experiments within the same county-domain block, we only count the first one. The regression controls for a full set of county $\times$ domain fixed effects, calendar year $\times$ domain fixed effects, and calendar year $\times$ county fixed effects. Specifically, we estimate the following regression: $y_{ikt} = \sum_m D^m_{ikt} \cdot \beta_m + \lambda_{it} + \delta_{kt} + \theta_{ik} + \varepsilon_{ikt}$, standard errors are clustered at the county level. The sample is balanced around the event time. We present the TWFE version as baseline, and also adjustments according to Sun and Abraham (2021) for robustness.

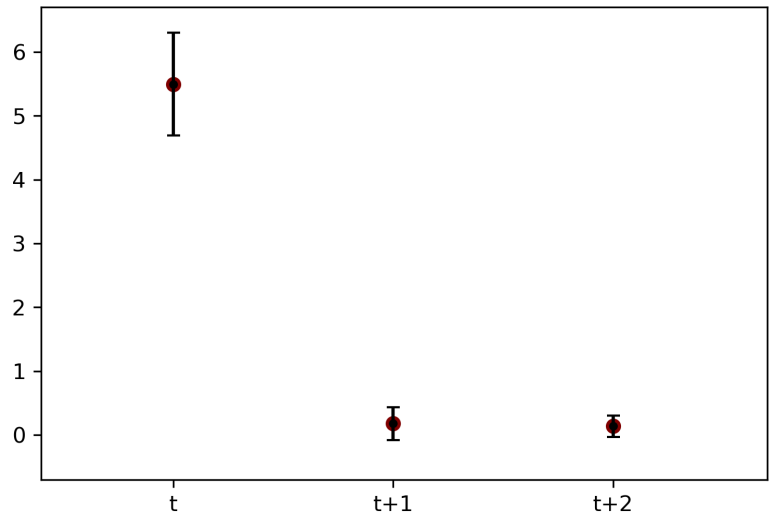


Figure A.13: This figure plots the point estimates and confidence intervals of iv_t, iv_{t+1}, and iv_{t+2}, in a unified first stage land revenue regression. We control for county and year fixed effects. Specifically, we estimate a set of $\hat{\beta}_i$s from a regression $Land\_revenue_{it} = \beta_1 iv_{i,t} + \beta_2 iv_{i,t+1} + \beta_3 iv_{i,t+3} + X'_{it}\beta + \delta_i + \gamma_t + \epsilon_{it}$. As we can see, iv_t is strongly correlated with land revenue$_i t$, while the lead terms of the IV have no predicting power on land revenue$_i t$. Standard errors are clustered at the county level.

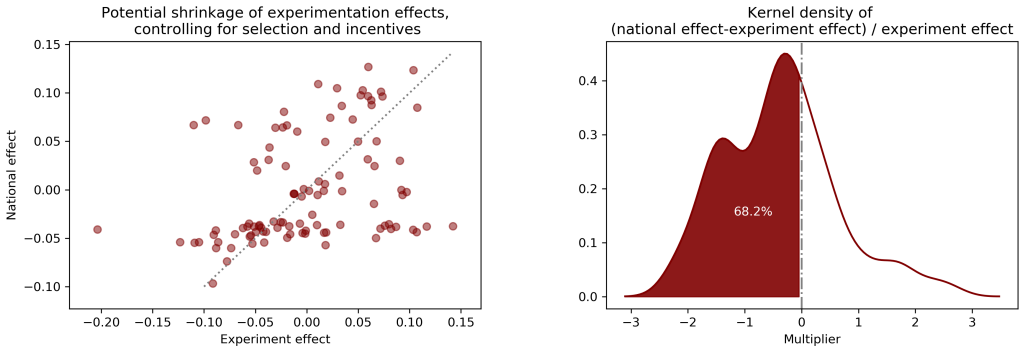


Figure A.14: These plots demonstrate how policy effects shrink between the experimentation and roll-out stages. In Panel A, we plot policy effect during national roll-out (y-axis) against experimentation effect of the same policy (x-axis), both residualized by the number of experiment sites. In Panel B, we compute the the difference between residualized policy effect during national roll-out and residualized policy effect during experimentation, and then took its ratio over the residualized policy effect during experimentation, and plot its distribution.

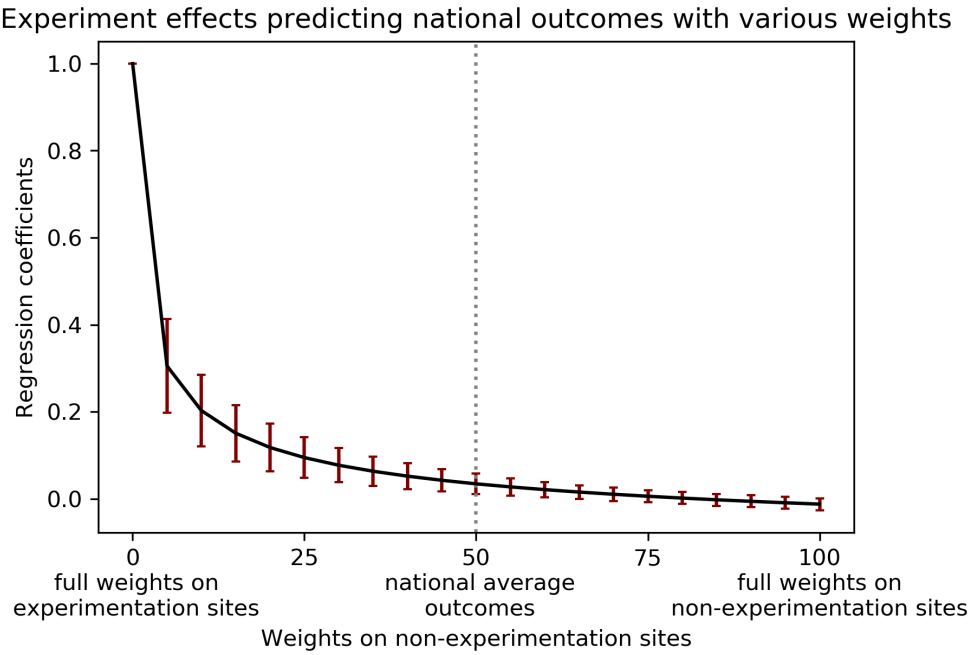
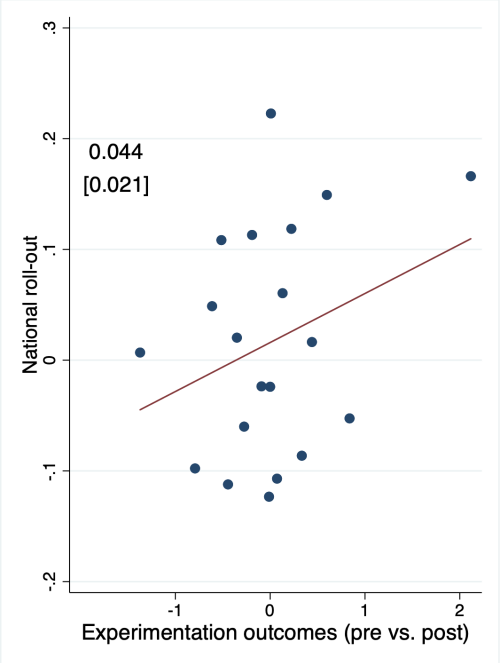
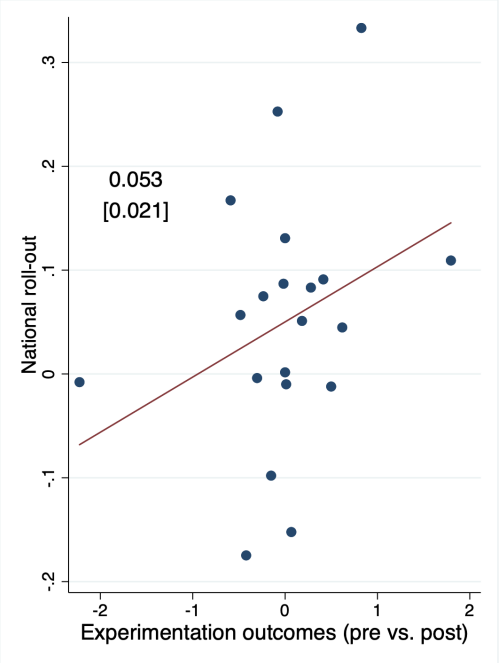


Figure A.15: On two ends of the spectrum are two extreme configurations: to the left, we place the entirety of the weight of the policies' national effects on those experimentation sites — thus, the predicted regression coefficients are 1, by construction. To the right end of the spectrum, we place the entirety of the weight of the policies' national effects on those non-experimentation sites. The mid-point of the spectrum represents equal weights across all localities in the country, and the x-axis represents continuous variations from one end of the spectrum to the other in the weighted national policy effects

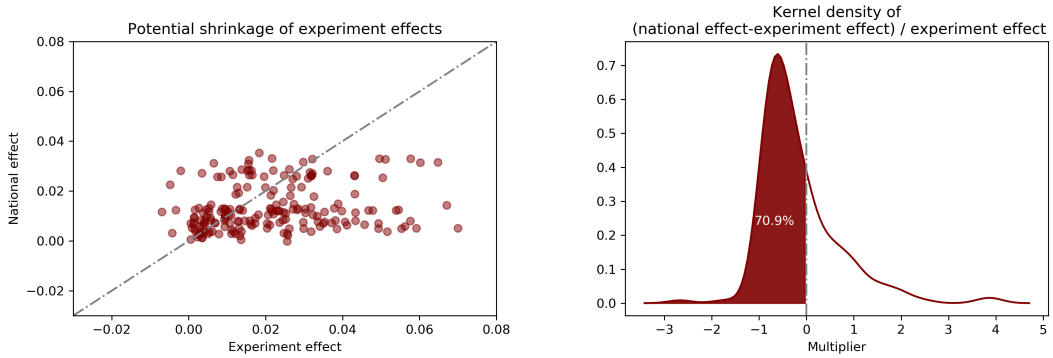


(a) Panel A. Fiscal revenue

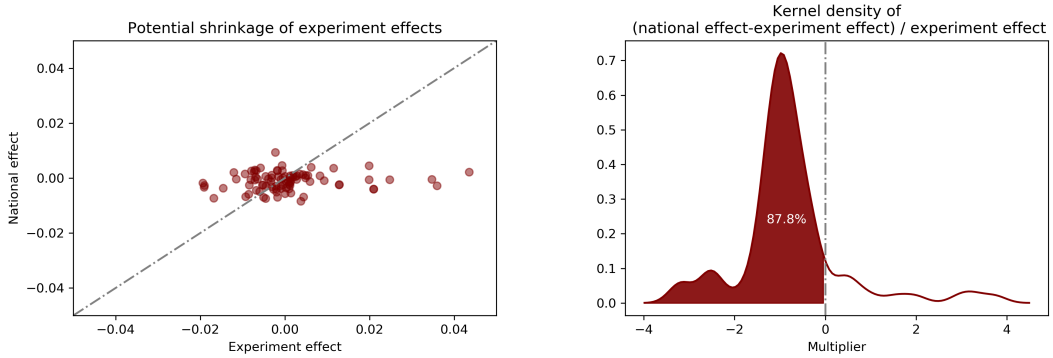


(b) Panel B. Night light luminosity

Figure A.16: This figure plots the correlations between naive ATE estimates and policy rollout parallel to Figure 4. Instead of computing average treatment effect with GDP per capita, which is prone to mis-reporting and manual inflation, we use two alternative measures that are harder to manipulate – fiscal revenue in Panel A, and nightlight luminosity index in Panel B.
The nightlight luminosity index is constructed following Martinez (2022), where we obtained satellite stable-light images from NOAA websites from 1992 to 2013 in TIFF format. In the raw images, each pixel corresponds to a 30 arc-second grid, with an approximate pixel size of 0.86 square kilometers at the equator, and it is coded as a discrete luminosity index from 0 to 65. If there are multiple satellites in the airspace, we take the average across images. We aggregate those indices to prefecture level in China by taking the average across prefecture shapefile boundaries. We then treat them as GDP per capita equivalents and follow the exact empirical specifications as we did in Figure 4, Panel A.



(a) Panel A. Fiscal revenue



(b) Panel B. Night light luminosity

Figure A.17: These figures plots the policy effect shrinkage between experimentation stage and roll-out stage using alternative measures: local fiscal revenue in Panel A, and nightlight luminosity index in Panel B. In Panel A, we plot policy effect during national roll-out (y-axis) against experimentation effect of the same policy (x-axis). In Panel B, we compute the the difference between policy effect during national roll-out and policy effect during experimentation, take its ratio over the experiment effect, and plot its distribution.
The nightlight luminosity index is constructed following Martinez (2022), where we obtained satellite stable-light images from NOAA websites from 1992 to 2013 in TIFF format. In the raw images, each pixel corresponds to a 30 arc-second grid, with an approximate pixel size of 0.86 square kilometers at the equator, and it is coded as a discrete luminosity index from 0 to 65. If there are multiple satellites in the airspace, we take the average across images. We aggregate those indices to prefecture level in China by taking the average across prefecture shapefile boundaries.

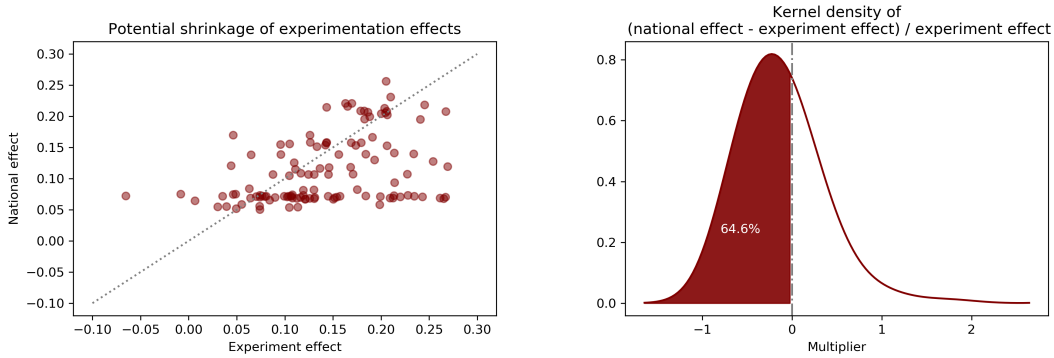


Figure A.18: National treatment effect deflation for policies targeting short-run effects. We define short-run policies by looking for keywords specifying specific date of evaluation or roll-out in all the experimentation-related documents. And that date has to be less than or equal to 3 years. The empirical specification follows exactly Figure 5

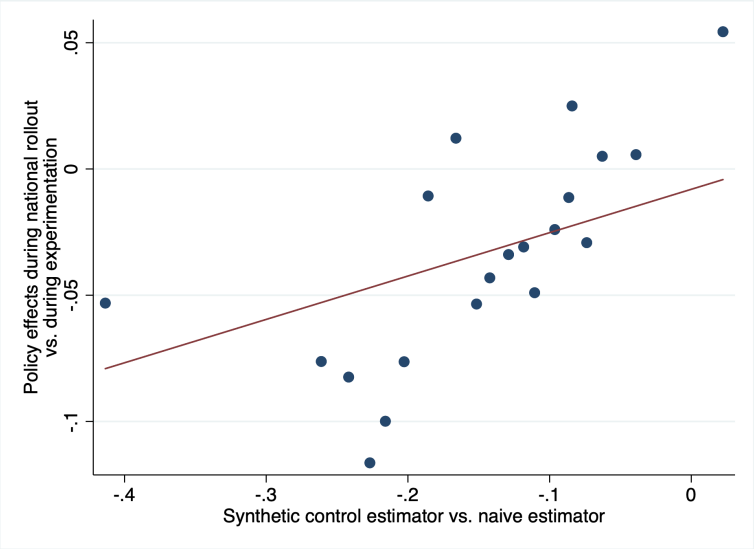


Figure A.19: This figure presents the correlation between the precision of estimation, and the policy effect deflation. On the x-axis, we compute the gap between synthetic control estimator and naïve estimator of experimentation effect; on the y-axis, we compute the gap between policy effect during national rollout and that during experiment effect. We retrieve our synthetic control estimator à la Xu (2017) matching on prefectural-level fiscal income, GDP per capita, and politician career incentive.

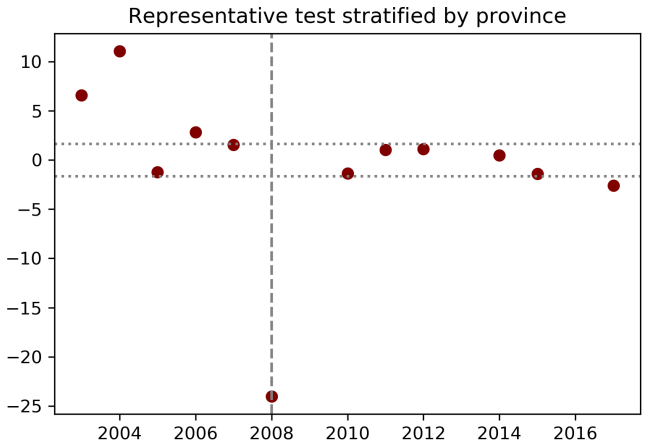
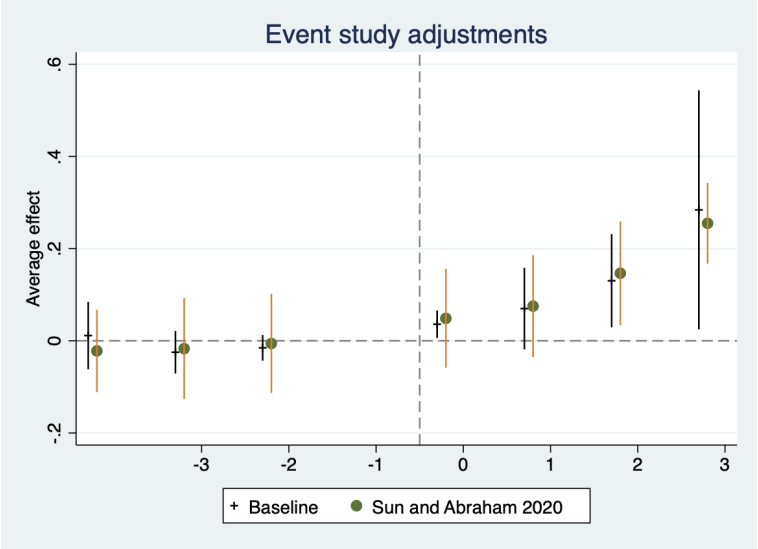
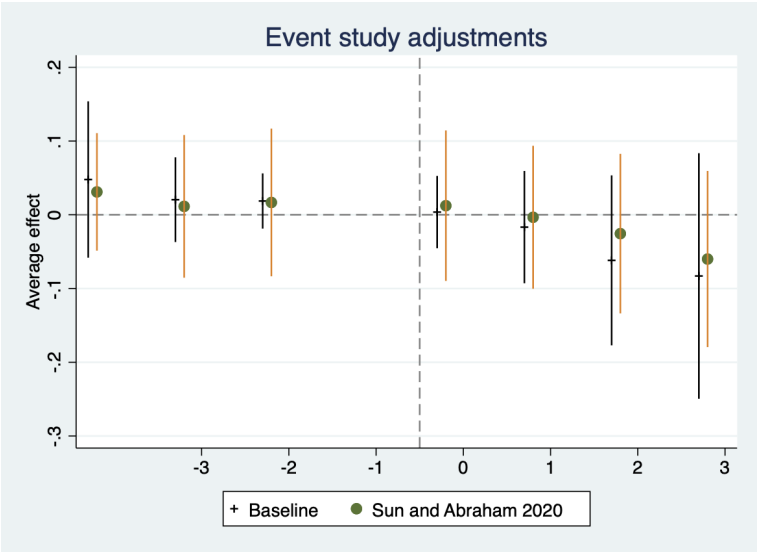


Figure A.20: This plot presents representative test of experimentation sites in the county fiscal empowerment reform. We conduct stratified Fisher randomization tests with student-t statistics and provincial strata. Within each province, we view counties that engage in the experimentation for the first time as units of the treatment group, the rest as control. Provincial level t-statistics are weighted and standard errors are estimated based on Miratrix, Sekhon, and Yu (2013). The grey horizontal lines indicate the asymptotic 95% confidence intervals within which representative assignment of experimentation sites cannot be rejected.



Panel A: Counties before 2007



Panel B: Counties after 2007

Figure A.21: This figure plots the effect of county fiscal empowerment reform on experimentation sites' local GDP per capita. Specifically, we estimate the following event study model for county c in year t: $y_{ct} = \sum_k D_{ct}^k \cdot \beta_k + \delta_c + \theta_t + \varepsilon_{ct}$, and report the coefficients for the subsamples of rich vs. poor counties. The standard errors are clustered at the province level. We control for a full set of county and calendar year fixed effects. The top panel plots the the effect among counties participated in the experiment before 2007, and bottom panel plots the effect among counties participated in the experiment after 2007.

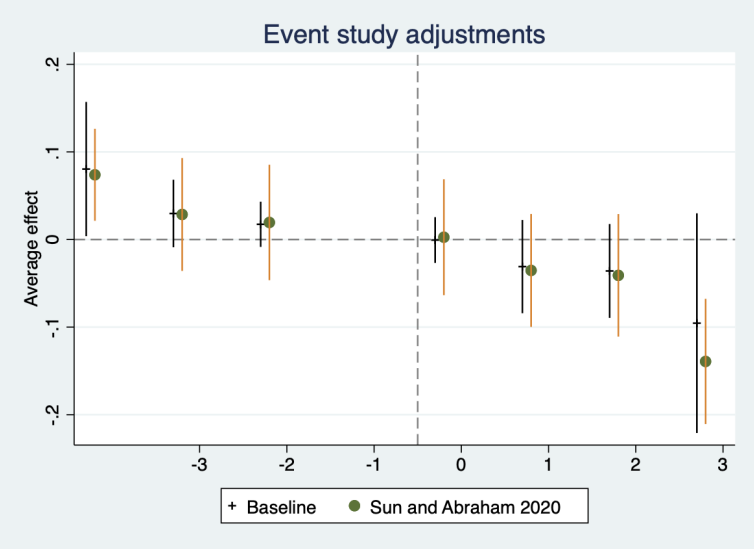


Figure A.22: This figure plots the effect of county fiscal empowerment reform on experimentation sites' local GDP per capita, among counties participated in the experiment before 2007 and whose pre-experimentation GDP per capita was below median. Specifically, we estimate the following event study model for county c in year t: $y_{ct} = \sum_k D_{ct}^k \cdot \beta_k + \delta_c + \theta_t + \varepsilon_{ct}$, and report the coefficients for the subsamples of rich vs. poor counties. The standard errors are clustered at the province level. We control for a full set of county and calendar year fixed effects.

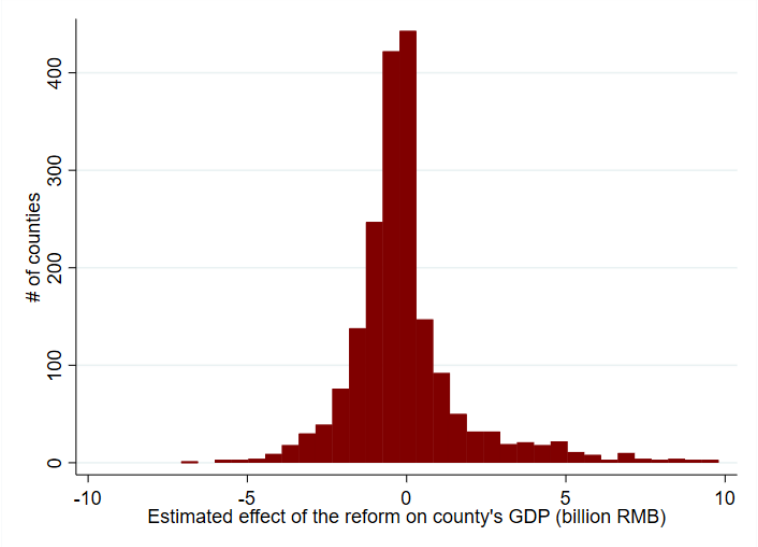


Figure A.23: This figure presents the distribution of simulated effect of county fiscal empowerment reform on local GDP per capita across Chinese counties. We extrapolate the estimated treatment effect among experimentation sites to all counties nationwide, allowing for the effect to be heterogeneous with respect to pre-reform local GDP per capita. Specifically, we first estimate the following model:
$y_{it} = \beta_1 Reform_{it} + \beta_2 Reform_{it} \times GDPPC_{i,2001} + \gamma_t + \alpha_i + \sigma_{t,prov} + \epsilon_{it}$, and then simulate the outcome assuming in the world where everyone is treated based on the heterogeneity of pre-period GDP per capita.

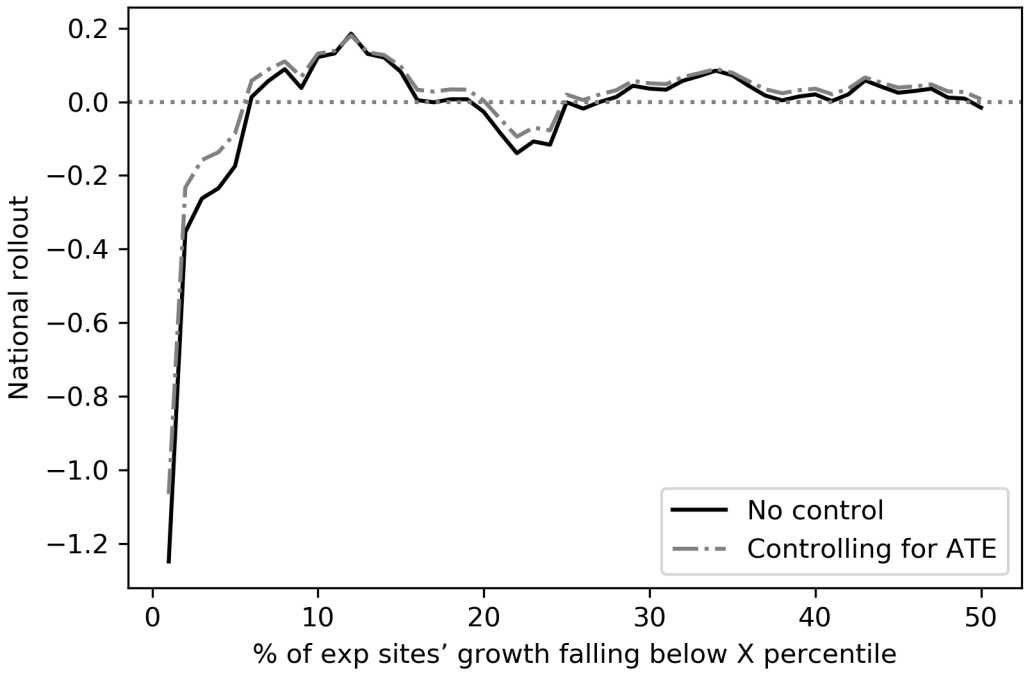


Figure A.24: In this figure we plot policy rollout against the presence of extremely bad-preforming localities. For each x on the axis, we regress national rollout on the % of experiment sites' GDP pc growth falling below x percentile nation-wide, and plot the regression coefficients together in one plot. As a robustness check, we control the average treatment effects in the regressions and plot the alternative set of coefficients in grey dashed lines.

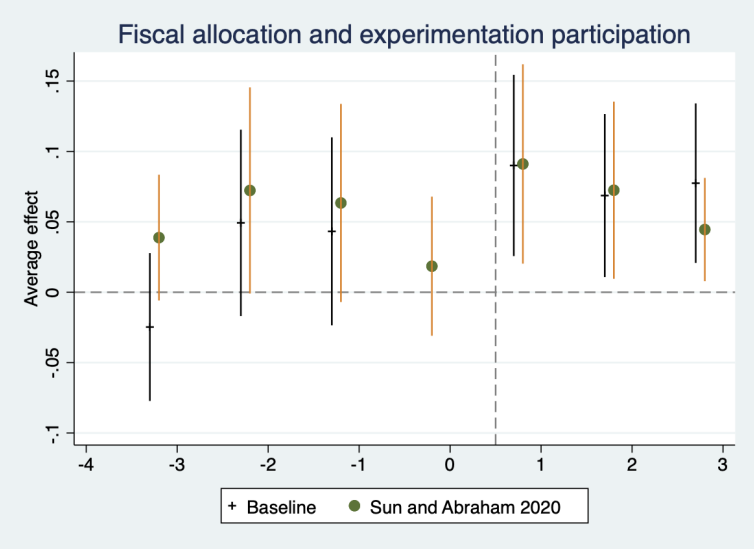


Figure A.25: This figure plots the event study estimates on a province's probability of being selected as an experimentation site after it becomes connected to a ministry due to political turnovers at the ministerial level. Specifically, we estimate the following econometric model using ministry-province-year level data: $y_{mpt} = \alpha \cdot Connection_{mpt} + \delta_{mp} + \theta_t + \varepsilon_{mpt}$, 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; θ_t is year fixed effects; and δ_{mp} stands for province-by-ministry fixed effects. Standard errors are clustered at the province $\times$ ministry level. All periods beyond the shown leads and lags are accumulated into final points

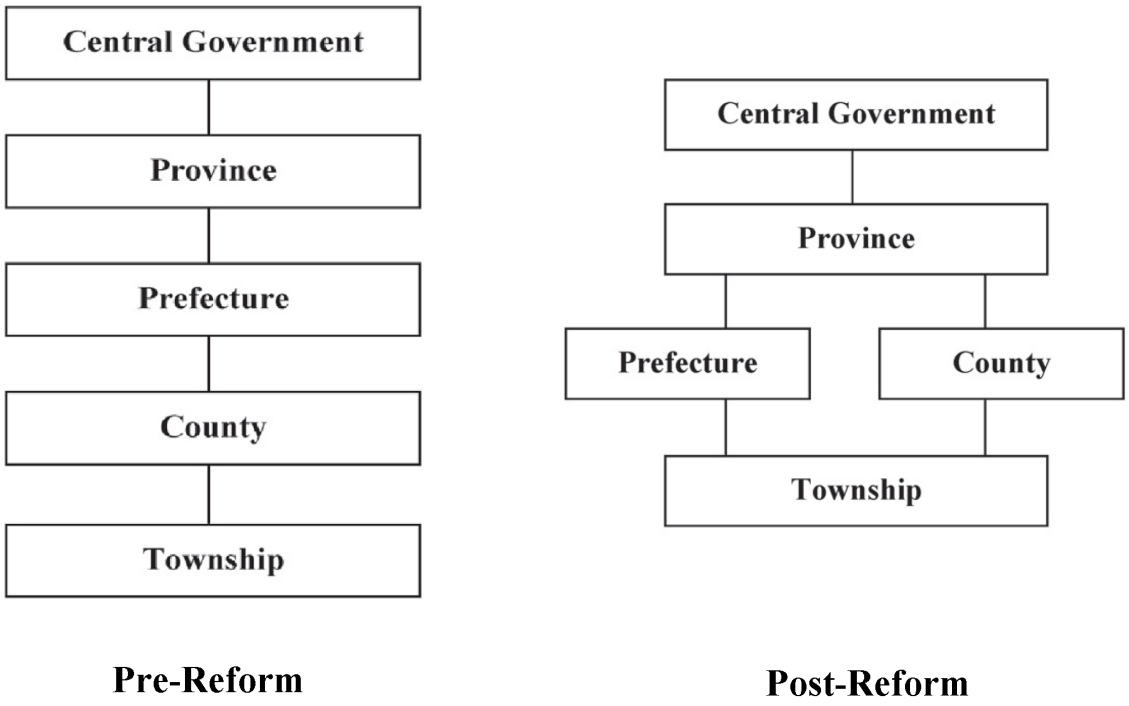


Figure A.26: Reproduced from Li, Lu, and Wang (2016). Illustration of county fiscal empowerment reform. After the reform, the provincial government could directly manage some of its counties, bypassing the prefectural cities, which grants county governments with more fiscal autonomy.

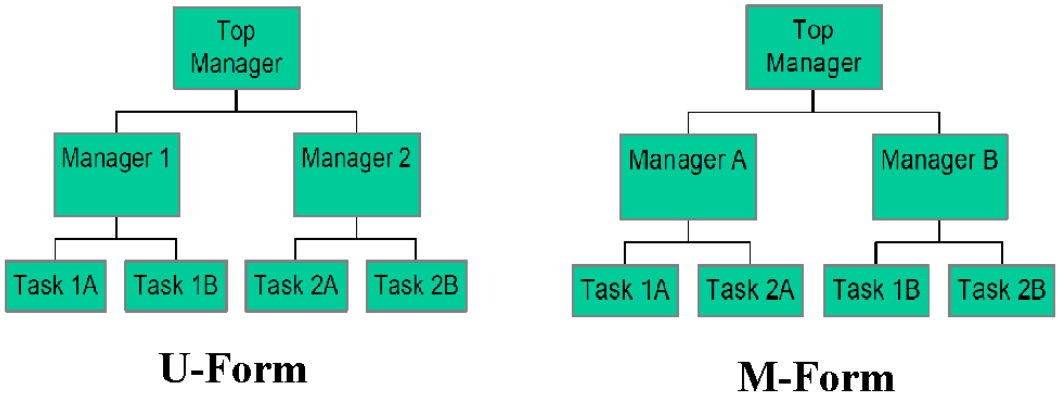


Figure A.27: Reproduced from Qian, Roland, and Xu (2006). Illustration of a shift from M-form to U-form. The top manager (ministers at central government branches in our case) have more administrative and personnel authority on its branches.

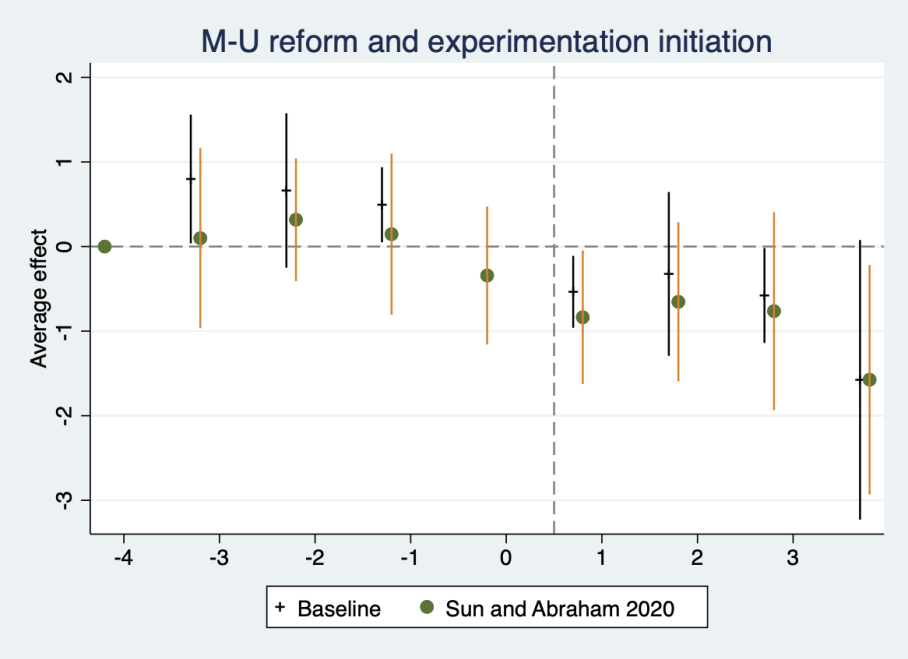


Figure A.28: This figure shows the decrease of policy experiments initiated by ministries as they transition from M-form into U-form. Specifically, we estimate the following event study model: $y_{mt} = \sum_k D^k_{mt} \cdot \beta_k + \delta_m + \theta_t + \varepsilon_{mt}$, where y_{mt} indicates the number of experiments initiated by ministry m in year t. X-axis indicates the time relative to the transition. The point estimates and confidence intervals are computed from a standard event study design, as well as adjustments recommended by Sun and Abraham (2021), controlling for ministry and calendar year fixed effects. Standard errors are clustered at the ministry level. All periods beyond the shown leads and lags are accumulated into final points.

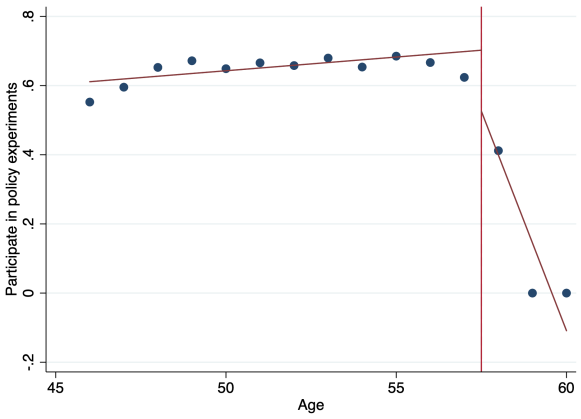


Figure A.29: In this figure, we plot the average number of experimentation participation against local politician's running age at prefecture level. Provincial policies are counted towards each of its prefectures, and county policies are counted once at the prefecture in which it resides. We add one to a politician's age if he or she is born on or after July.

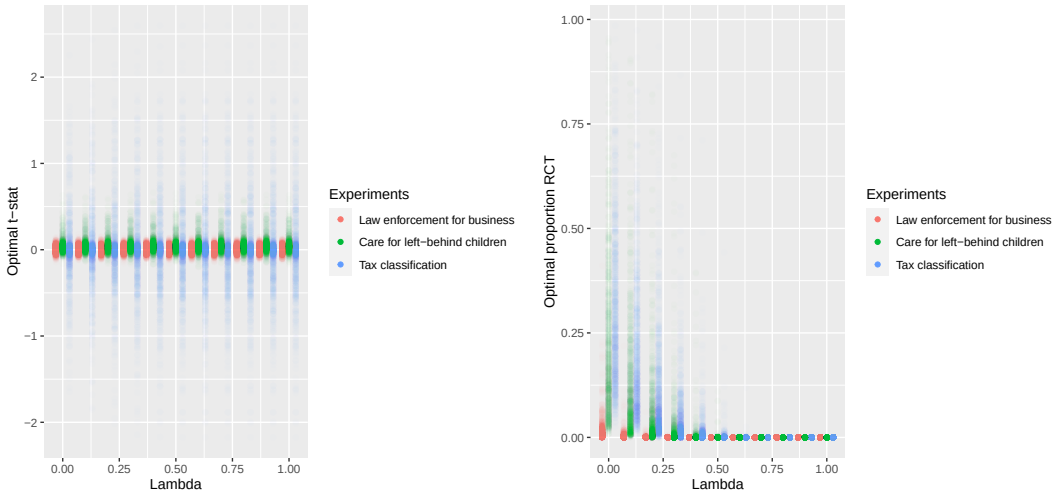


Figure A.30: This plot shows optimal t-statistics (left) and RCT vs. deterministic experimental policy (right) for simulations calibrated using three different policy experiments conducted in China following the model in Banerjee et al. 2020. Lambda ranges from 1 (full weight on decision maker's utility) 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.

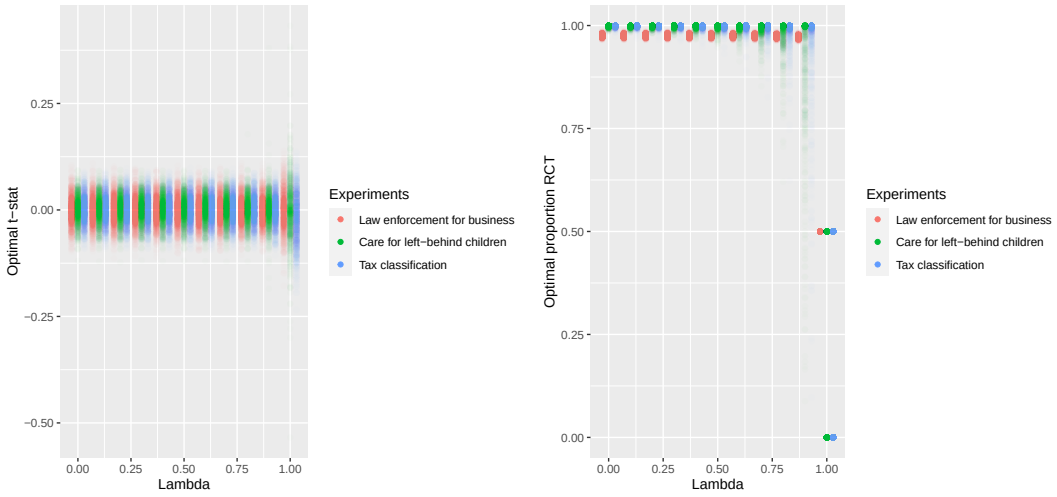


Figure A.31: This plot shows optimal t-statistics (left) and RCT vs. deterministic experimental policy (right) for simulations calibrated using three different policy experiments conducted in China following the model in Banerjee et al. 2020 with differential quality of information. Lambda ranges from 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.

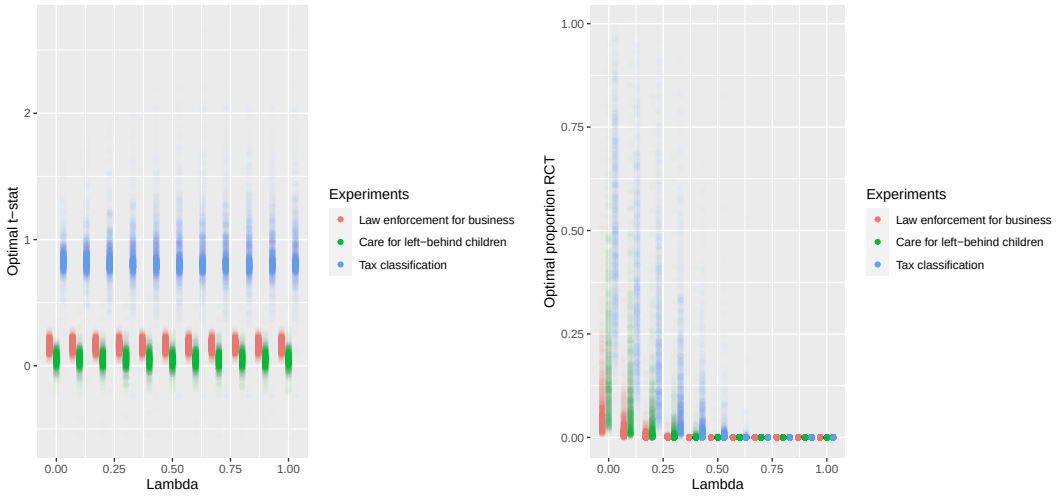


Figure A.32: This plot shows optimal t-statistics (left) and RCT vs. deterministic experimental policy (right) for simulations calibrated using three different policy experiments conducted in China following the model in Banerjee et al. 2020 with subject consent. Lambda ranges from 1 (full weight on 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.

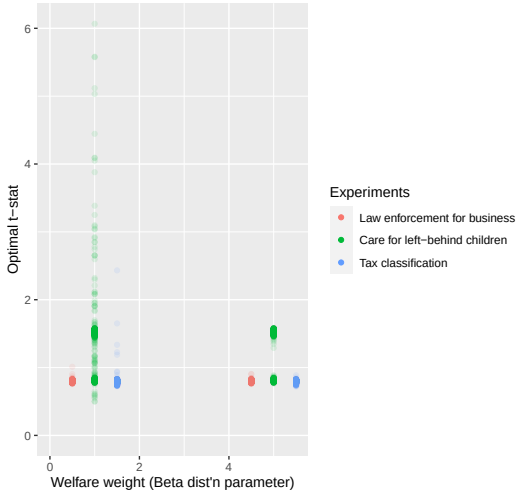


Figure A.33: This plot shows optimal t-statistics for simulations calibrated using three different policy experiments conducted in China following the model in Narita 2021. The welfare weight δ corresponds to a parameterization of the Beta distribution Beta$(\delta, 10 - \delta)$ where higher values of δ place more weight on wealthier counties by GDP per capita.

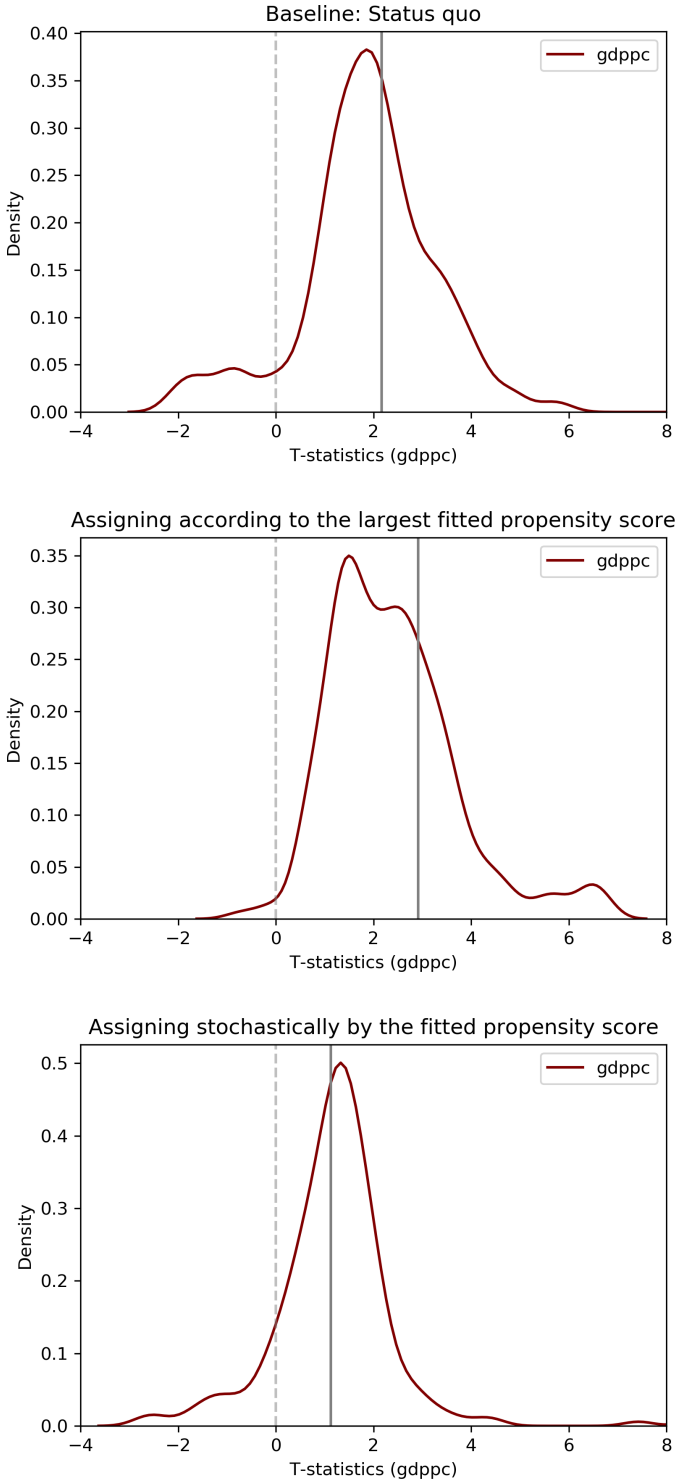


Figure A.34: This figure presents additional robustness exercises of the representativeness tests' t-statistics distribution, using the same test procedures as Figure 3. In Panel A, we only include policy experiments targeting prefectures and we exclude the 4 provincial-level municipal units from the sample. Panel B shows the simulated results if experimentation sites are assigned sequentially to the prefectural units with the largest fitted propensity score; where the propensity score is estimated as the the probability of each prefecture being assigned to a particular treatment given a set of observed characteristics (politicians' career incentive, political hierarchical level, presence of political patronage). In Panel C, we simulate the site-selection process with the fitted propensity score as our weights. We do 500 random draws with replacement and plotted the distribution of *t*-statistics. The gray vertical line indicates the mean of t-stats.

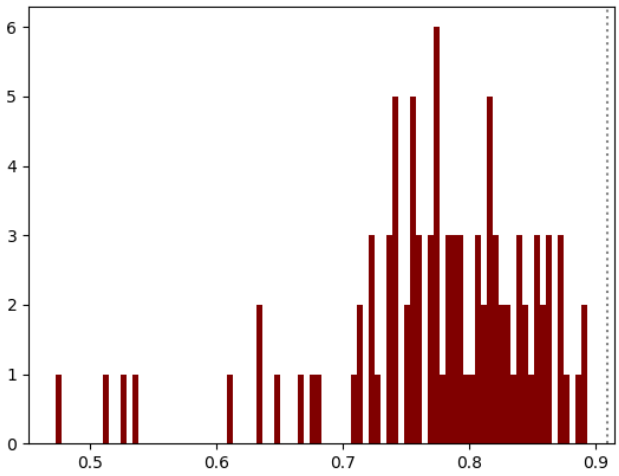


Figure A.35: This figure plots the distribution of pairwise similarity index between a random 100 sample of policy experimentation documents and the Shenzhen 2014 Carbon Emission Tax (CET) regulations. The grey dashed line presents pairwise similarity index for the comparison between the 2014 version of CET regulations in Shenzhen, and the 2022 version that builds on their ancestors, which is a priori the most similar document

Table A.1: Examples for policy experimentation in China

Category	Policy experiment	Summary	Document
Transportation	Low carbon-emission transportation system	Promoting and subsidizing energy-saving and low-carbon means of transportation	link
	Vocational education for bus drivers	Setting up specialized schools for vocational bus drivers and lifting the age limit for those drivers	link
	Integration of big data in transportation	A comprehensive digitization reform on transportation data collection, analysis, and regulation	link
Population and Health	Hierarchical medical system	Encouraging different medical institutions to undertake the treatment of different diseases graded according to the priority and difficulty of treatment	link
	New rural cooperative medical system	Providing heavily subsidized medical insurance for rural residents and monitoring construction of village hospitals	link
	Monitoring and reporting of adverse drug reactions	Streamlining the process of adverse drug reactions monitoring and reporting	link
Media	Digital cable broadcasting and television services	Starting paid channels of digital TV and promoting digital TV set-up box nationwide	link
	Disclosure of annual reports for printing enterprises	Mandatory disclosure of annual reports for all printing companies that have received a printing business license (photocopying excluded)	link
	Online publishing supervision system	A prototype to conduct internet content supervision and censorship	link
Internal Affairs	Elderly-friendly communities and cities	Promotion of senior-friendly facilities, designs and public goods	link
	Comprehensive reform of social assistance	Exploring the mechanism for quantifying and determining social assistance standards and methods for identifying social assistance recipients	link
	Poverty alleviation through technology	Promoting and subsidizing technology-based poverty alleviation programs	link
Agriculture	Agricultural standardization	Setting standards for agricultural products and their primary processing products in terms of the variety, specifications, quality, grade and safety, health requirements	link
	Crop rotation	In areas with low crop production efficiency and obvious ecological degradation in autumn and winter, promoting crop rotation and crop exchange, winter tillage, fallow cultivation and fertilization, etc.	link

	Registration of rural land contract management rights	Standardizing the registration and certification of land contracting rights and land management rights	link
Labor and Personnel	Reform of salary system in public hospitals	Granting hospitals the degree of freedom to decide the salary of the doctors, experimenting incentive pay	link
	Reform of personnel system in public institutions	Allowing dismissals, canceling administrative hierarchy in public institutions, promoting flexible pay system	link
	Work Injury Prevention	Improving and creating working conditions conducive to safety and health, reducing work-related accidents and the hidden danger of occupational diseases	link
Development and Reform	Green finance	Supporting economic activities for environmental improvement, climate change response and resource conservation and efficient use	link
	Key development and opening up experimental zone	Subsidizing medicare, infrastructure construction, and poverty alleviation programs of residents at the border area	link
	Carbon emission rights trading	Defining the greenhouse gas cap on the emissions allowed in a well-defined sector (coverage) of the economy, when emission permits or allowances are issued or sold (allocated) to entities that are included in the carbon market	link
Judicial Supervision	Community correction	The execution of non-custodial sentences in which eligible offenders are placed in the community, where special state agencies correct their criminal psychology and behavioral vices and facilitate their smooth return to society	link
	Jury system reform	Introducing the jury system and open courtroom reform, promoting randomized jury invitation	link
	Attorney mediation	As a new dispute resolution mechanism, giving full play to the role of lawyers to establish a lawyer mediation work model	link
Commerce and Trade	Recycling system for renewable resources	Setting up standardized recycling sites in communities, and entering designated markets for standardized trading and centralized processing of renewable resources	link
	SOE debt reduction and relief project	Canceling the debt of State-owned enterprises due to government plans to transfer, price changes and other reasons in the planned economy	link

	Retention of profits in state-owned industrial enterprises	Changing the original provision of the full profit retention method to the base profit retention plus growth profit retention method	link
Industry and IT	Online monitoring of industrial energy consumption	Establishing energy consumption real-time monitoring system for large coal-burning enterprises in the monitoring scope	link
	Regional brand building for industrial clusters	Subsidizing and Curating national brand in large industrial clusters	link
	Credit system for express industry	Centralize social credit file management system and experiment rules of credit rating for service providers of the express industry	link
Market Supervision	Patent insurance	The insurance company compensating the insured for the investigation costs and legal fees incurred by the insured for patent defense in accordance with the contract	link
	Trademark agency system	The agent engaging in civil legal acts in the name of the represented party, and the legal consequences arising therefrom being directly attributed to the represented party	link
	Disclosure of consumer complaints	Mandatory disclosure of all consumer complaints or disputes in a centralized platform	link
Education, Science, Culture and Sports	Sports test for high school entrance exams	Introducing a mandatory sports test to the standardized test for high school admission	link
	Reform of college English teaching	Standardizing the teaching materials and introducing CET-4 as a prerequisite for college graduation	link
	Primary and secondary school teacher qualification examination	Introducing a qualification test for teachers before assuming a job, and asking for periodic registration	link
State Council	Shareholding system with public offering	Promoting a fundamental ownership reform, establishing the stock market, and allowing shares to be traded among the public	link
	Separation of license and permits	Simplifying the company registration process via separating the applications of permits, which is fast, and licenses, which is demanding	link
	Rural tax and fee reform	Completely abolishing agricultural tax	link
	Government procurement credit guarantee	Guarantees provided by professional guarantee organizations to the government procurers on behalf of suppliers	link

Finance and Taxation	Integration and use of fiscal agricultural funds	For counties in poverty, centralize the allocation of fiscal transfers for agricultural purposes	link
	VAT reform	Collection of value-added tax instead of sales tax among all industries	link
Resources, Energy, and Environment	Rehabilitation of Rural Dangerous Houses	Identification and reconstruction of dilapidated houses, after 2008 Wenchuan Earthquake	link
	Environmental supervision during construction period	Establishing a dedicated team for environmental monitoring during construction	link
	Electricity trading	Direct trading of electricity between power plants and companies, without accessing the national grid	link
Finance	Futures exchange	Establishing a centralize market for future trading	link
	Management of banks' asset-liability ratio	Impose mandatory minimum rate of asset-liability ratio in commercial banks and rural credit unions	link
	Loans for rural small business	Setting up and subsidizing loan options for small business in rural area	link

Table A.2: Comprehensiveness checks for the *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 on the website (e.g., Fiscal and Tax; Natural Resources). We make sure that the numbers being compared come from the same time frame.

Table A.3: Positive selection, ex-ante uncertainty, complexity

	Regression coef.	s.e.
Panel A: Ex-ante certainty		
Ex-ante rollout schedule	-0.215	0.256
Scheduled evaluation date	-1.301**	0.517
Academic consensus index	-0.880**	0.393
Certainty index	-1.550**	0.683
Panel B: Complexity		
Multi-ministry cooperation	0.130	0.230
Word count of first policy document	0.787***	0.162
Number of relevant policy documents	0.391**	0.182
Word count of all relevant documents	0.432**	0.191
Average word count of policy documents	0.857***	0.189
Duration of policy experiment	0.322	0.327
Number of local government follow-up policies	0.229	0.223
Complexity index	1.290***	0.381
Panel C: Administrative level		
County or prefecture level	5.449***	0.351

Notes: In this table we regress the level of positive selection on a spectrum of indices measuring ex-ante policy (un)certainty and complexity. Specifically, we report the estimated coefficients and robust standard errors from the model $y_i = \beta X_i + \delta_m + \theta_t + \varepsilon_i mt$, where y_i is the t-statistics for policy i comparing fiscal income between treatment and control localities. Ministry fixed effects (δ_m) and year fixed effects (θ_t) are controlled for across all regressions. We standardize all independent variables to zero mean and unit variance for interpretation purposes.

We measure the ex-ante certainty and complexity of policy experiments along various dimensions. Reading from the policy documents, we labeled (1) the presence of ex-ante rollout schedule in the first policy document and (2) the pre-scheduled date of policy evaluation. (3) For the academic consensus index, we first match policy keywords to all the academic papers published between 2005 and 2017 by authors acknowledging at least one National Social Science Fund (NSSF) from the government. NSSF is known to be the most authoritative funds in the Chinese academia that is awarded to a large number of scholars. We then compare corresponding papers in a pair-wise way, calculating TF-IDF (term frequency–inverse document frequency) index for each pair, and took the median as final consensual index for the policy.

In panel B, we measure complexity by (1) an indicator variable of whether more than 1 ministry is involved in the experimentation; (2) the length of the document initiating the policy experiment; (3) the number of all relevant documents; (3) the total length of all relevant documents; (4) the average length of all relevant documents; (5) the actual duration of the policy experiment; and (6) the number of documents followed-up by the local government to echo the spirit or provide implementation details.

In the last row of Panel A (B), we constructed a certainty (complexity) index that is the arithmetic mean of all the standardized indices above in the same panel.

In Panel C, we provide a separate regression comparing the level of positive selection between province level experiments vs. county or prefecture level experiments.

Table A.4: Changes in positive experimentation sites selection over time

	Year		
	coef.	s.e.	coef / mean
	(1)	(2)	(3)
Panel A: Full sample			
OLS	-0.067	0.038	-0.013
Ministry FE	-0.113	0.062	-0.023
Panel B: By ministry			
Transportation	-0.294	0.096	-0.093
Agriculture	-0.305	0.139	-0.080
Law	-0.379	0.216	-0.078
Development and reform	-0.259	0.242	-0.069
Commerce	-0.174	0.120	-0.028
Education	-0.136	0.107	-0.027
Industry and information technology	-0.181	0.214	-0.026
Labor	-0.071	0.194	-0.015
Tax	-0.080	0.113	-0.015
Population and healthcare	-0.056	0.124	-0.012
Market supervision	0.027	0.126	0.005
Resource, energy & environment	0.097	0.078	0.026
Finance	0.253	0.305	0.030
Domestic affairs	0.201	0.161	0.052
State ministry	0.540	0.268	0.100
Media	0.280	0.000	0.630

Notes: In this table we regress the level of positive selection on calendar year. Each row is a separate regression. Specifically, we report the estimated coefficients and robust standard errors from the model $y_i = \beta t_i + \delta_m + \varepsilon_i$, where y_i is the t-statistics comparing fiscal income between treatment and control localities (mean = 5.01, s.d.=5.22), and t_i is 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. Ministries are sorted in ascending order by column 3. In the second row of panel A, we also include a specification where we control for ministry fixed effects.

Table A.5: Politician promotion and experimentation participation

	Promotion					
	(1)	(2)	(3)	(4)	(5)	(6)
Participated in any experiments	0.057	0.017				
	(0.044)	(0.046)				
... successful experiments			0.109***	0.094**		
			(0.037)	(0.042)		
... small-scale & successful experiments					0.139**	0.121*
					(0.066)	(0.066)
# of obs.	1,052	1,052	1,052	1,052	1,052	1,052
# of clusters	316	316	316	316	316	316
Mean of DV	0.423	0.423	0.423	0.423	0.423	0.423
Prefecture FE	Yes	Yes	Yes	Yes	Yes	Yes
Tenure FE	No	Yes	No	Yes	No	Yes

Note: Using prefecture-term level data, we estimate the following model for the party secretary of prefecture p during his term t: $Promotion_{pt} = \alpha \cdot Exp_{pt} + \delta_p + \gamma_t + \epsilon_{pt}$. Standard errors clustered at the prefecture level are reported below the estimates. Promotion follows the canonical definition à la Wang, Zhang, and Zhou (2020). 77.4% of politicians participated in at least one experiment during his or her term. We define successful experiments as those experiments rolled-out eventually to the entire country. 65.5% of politicians participated in at least one successful experiment during tenure. We define small-scale experiments as policies that are trialed in less than 10 localities across all waves. 4.59% of politicians participated in small-scale experiments.

Table A.6: Local fiscal expenditure during policy experimentation

	Share of fiscal expenditure on experimentation-related domains			
	(1)	(2)	(3)	(4)
# of experiments	0.003***	0.002***	0.002***	0.003***
	(0.001)	(0.000)	(0.000)	(0.001)
# of experiments $\times$**1** {over 58}	-0.004**	-0.002**	-0.002*	-0.002**
	(0.002)	(0.001)	(0.001)	(0.001)
# of obs.	150,977	150,977	150,977	97,700
# of clusters	1,973	1,973	1,973	1,894
Mean of DV	0.173	0.173	0.173	0.174
County by category FE	No	Yes	Yes	Yes
Year by county FE	Yes	No	Yes	Yes
Category by year FE	Yes	Yes	Yes	Yes
Sample	Full sample	Full sample	Full sample	Age $>$ 50

Notes: In this table, we explore the heterogeneity of fiscal allocation with respect to politicians' career incentives. Specifically, we estimate the following model: $y_{ikt} = \alpha \cdot Exp_{ikt} + \beta \cdot Exp_{ikt} \times \mathbf{1}\{over58_{it}\} + \lambda_{it} + \delta_{kt} + \theta_{ik} + \varepsilon_{ikt}$. Standard errors clustered at the county level are reported in the parentheses. Career incentives are measured by an indicator variable showing whether the politician's age goes above 58. The mean of this indicator variable is 0.106. Constrained with the availability of fiscal expenditure data, we focus our analysis on policy experiments that happen between 1993 and 2006. In column 4 we focus on all county-year-domain grids where the responsible politician is over 50 years old, thereby effectively excluding very-young politicians to focus on the local effect.

Table A.7: Local fiscal expenditure during policy experimentation

	Share of fiscal expenditure on experimentation-related domains		
	(1)	(2)	(3)
# of small-scale experiments	-0.026***	-0.006***	-0.007**
	(0.005)	(0.002)	(0.003)
# × career incentives	0.054***	0.014***	0.017***
	(0.010)	(0.005)	(0.006)
# of obs.	142,128	142,116	142,116
# of clusters	1973	1973	1973
Mean of DV	0.173	0.173	0.173
County by category FE	No	Yes	Yes
Year by county FE	Yes	No	Yes
Category by year FE	Yes	Yes	Yes

Notes: Standard errors clustered at the county level are reported in the parentheses. This table follows our main triple difference specification where we estimate the following model: $y_{ikt} = \alpha \cdot Exp_{ikt} + Exp_{ikt} \times \mathbf{1}\{Incentives_{it}\} + \lambda_{it} + \delta_{kt} + \theta_{ik} + \varepsilon_{ikt}$. However, Exp_{ikt} in this table only counted smaller-scale experiments (policies with below average number of sites, mean=0.099, s.d.=0.358). Career incentives are measured as the ex-ante probability of promotion projected by the start age of tenure and hierarchical level. (mean=0.481, s.d.=0.075) We find larger politician efforts on small-scale experiments. Constrained with the availability of fiscal expenditure data, we focus our analysis on policy experiments that happen between 1993 and 2006.

Table A.8: Local fiscal expenditure during policy experimentation

	Share of fiscal expenditure on experimentation-related domains		
	(1)	(2)	(3)
# of experiments	0.0003	0.001***	0.002***
	(0.001)	(0.0004)	(0.001)
# of experiments × **1**{only experiment in domain-year}	0.010***	0.001**	0.001**
	(0.001)	(0.001)	(0.001)
# of obs.	142,128	142,116	142,116
# of clusters	1,973	1,973	1,973
Mean of DV	0.173	0.173	0.173
County by category FE	No	Yes	Yes
Year by county FE	Yes	No	Yes
Category by year FE	Yes	Yes	Yes

Notes: In this table, we explore the heterogeneity of fiscal allocation with respect to political attention. Specifically, we estimate the following model: $y_{ikt} = \alpha \cdot Exp_{ikt} + \beta \cdot Exp_{ikt} \times \mathbf{1}\{only\_experiment_i kt\} + \lambda_{it} + \delta_{kt} + \theta_{ik} + \varepsilon_{ikt}$. Standard errors clustered at the county level are reported below the estimates. $\mathbf{1}\{only\_experiment\}$ is an indicator showing whether the experiment is the only policy the local politician is monitoring in the specific year of the specific policy domain (mean=0.138). Constrained with the availability of fiscal expenditure data, we focus our analysis on policy experiments that happen between 1993 and 2006.

Table A.9: Fiscal allocation heterogeneity by localities' socioeconomic conditions

	Share of fiscal expenditure		
	(1)	(2)	(3)
Panel A: Interacting with GDP per capita			
# of experiments	0.0018*	0.0015***	0.0017***
	(0.0010)	(0.0005)	(0.0006)
# × GDP per capita	-0.00002	0.00003	0.0003
	(0.0005)	(0.0002)	(0.0003)
Panel B: Interacting with GDP			
# of experiments	0.002**	0.002***	0.002***
	(0.0008)	(0.0004)	(0.0005)
# × GDP	-0.0003	-0.0001	0.0001
	(0.0005)	(0.0002)	(0.0003)
# of obs.	102,197	102,197	102,197
# of clusters	1778	1778	1778
Mean of DV	0.173	0.173	0.173
County × category FE	No	Yes	Yes
Category × year FE	Yes	Yes	Yes
Year × County FE	Yes	No	Yes

Note: In this table, we explore the heterogeneity of fiscal allocation with respect to local socio-economic conditions. Specifically, we estimate the following model: $y_{ikt} = \alpha \cdot Exp_{ikt} + \beta \cdot Exp_{ikt} \times w_{it} + \lambda_{it} + \delta_{kt} + \theta_{ik} + \varepsilon_{ikt}$. Standard errors clustered at the county level are reported below the estimates. In this table, we regress total fiscal expenditure on the number of participated experiments. In addition, we interact the number of experiments with pre-experimentation socioeconomic conditions: Panel A focuses on localities' GDP per capita, and Panel B on total local GDP. In both panels, we standardize GDP (per capita) to zero mean and unit variance. Sample: policy experiments that happen between 1993 and 2006.

Table A.10: Local fiscal allocation for experiments with explicit central government fiscal support

	Share of fiscal expenditure					
	(1)	(2)	(3)	(4)	(5)	(6)
# of experiments	0.01876***	0.00304***	0.00390***	-0.00514	-0.00284*	-0.00302
	(0.00179)	(0.00061)	(0.00074)	(0.00412)	(0.00164)	(0.00199)
# × career incentive				0.05078***	0.01202***	0.01410***
				(0.00834)	(0.00335)	(0.00404)
# of obs.	142,116	142,116	142,116	142,116	142,116	142,116
# of clusters	1973	1973	1973	1973	1973	1973
Mean of DV	0.174	0.174	0.174	0.174	0.174	0.174
County × category FE	No	Yes	Yes	No	Yes	Yes
Category × year FE	Yes	Yes	Yes	Yes	Yes	Yes
Year × County FE	Yes	No	Yes	Yes	No	Yes

Notes: In this table, we re-estimate our baseline regression on a subsample of policy experiments with explicit central government fiscal support (as stated in the first and main central policy document, mean=0.096). Specifically, we estimate the following model: $y_{ikt} = \alpha \cdot Exp_{ikt} + \lambda_{it} + \delta_{kt} + \theta_{ik} + \varepsilon_{ikt}$. Standard errors clustered at the county level are reported below the estimates.

Table A.11: Endogenous efforts during experimentation: Policy-by-county level regression

	Share of fiscal expenditure on experimentation-related domains			
	(1)	(2)	(3)	(4)
Participated in experimentation	0.00370***	0.00344***	-0.0146***	-0.00244
	(0.000808)	(0.000693)	(0.00405)	(0.00337)
Participation × career incentive			0.0394***	0.0125*
			(0.00851)	(0.00716)
# of obs.	96,221	93,612	92,089	89,547
# of clusters	1885	1885	1880	1880
Mean of DV	0.185	0.185	0.185	0.185
Policy FE	Yes	Yes	Yes	Yes
County FE	Yes	No	Yes	No
County × domain FE	No	Yes	No	Yes

Notes: Standard errors clustered at the county level are reported in the parentheses. Here we present a specification where we run the regression at policy-by-county level. Specifically, we estimate the following model: $y_{ik} = \alpha \cdot Exp_{ip} + \lambda_i + \delta_p + \theta_{ik(p)} + \varepsilon_{ipk}$. The dependent variable is the share of experiment-related fiscal expenditure at the first year of the experiment, and the independent variable is an indicator of counties' experiment participation status, as well as its interaction with the career incentive of the prefecture mayor at that year. By doing this we are able to control for policy experiment fixed effects. Career incentives are measured as the ex-ante probability of promotion projected by the start age of tenure and hierarchical level (mean=0.481, s.d.=0.075). In columns 1 and 3 we additionally controlled for county fixed effects, and in columns 2 and 4 we additionally controlled for county × policy domain fixed effects.

Table A.12: Local fiscal expenditure during policy experimentation

	Share of fiscal expenditure on experimentation-related domains					
	(1)	(2)	(3)	(4)	(5)	(6)
# experiments	0.003***	0.002***	0.002***	-0.011***	-0.002*	-0.003
	(0.001)	(0.000)	(0.000)	(0.003)	(0.001)	(0.002)
# × career incentive				0.029***	0.008***	0.011***
				(0.006)	(0.003)	(0.003)
# concluded exp's	-0.0003	0.001	0.001	-0.028***	0.001	0.002
	(0.001)	(0.001)	(0.001)	(0.005)	(0.002)	(0.003)
# concluded × incentive				0.061***	-0.001	-0.003
				(0.010)	(0.005)	(0.007)
# of obs.	150,977	150,977	150,977	142,128	142,116	142,116
# of clusters	1973	1973	1973	1973	1973	1973
Mean of DV	0.173	0.173	0.173	0.173	0.173	0.173
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

Notes: Standard errors clustered at the county level are reported below the estimates. In this table, we horse-race the number of on-going experiments with the number of concluded experiments at the same year. Specifically, we estimate the following model: $y_{ikt} = \alpha \cdot Exp_{ikt} + \beta \cdot Concluded\_exp_{ikt} + \lambda_{it} + \delta_{kt} + \theta_{ik} + \varepsilon_{ikt}$. "Concluded exp" counts the number experiments that had been wrapped up 2 years ago. Career incentives are measured as the ex-ante probability of promotion projected by the start age of tenure and hierarchical level (mean=0.481, s.d.=0.075). We focus our analysis on policy experiments that happen between 1993 and 2006.

Table A.13: Strategic differentiation in experimentation plans

	Similarity index in experimentation plans			
	(1)	(2)	(3)	(4)
Career incentive	-0.057***	-0.056**	-0.057**	-0.055**
	(0.021)	(0.022)	(0.022)	(0.022)
# of obs.	3,970	3,970	3,970	3,970
# of clusters	233	233	233	233
Mean of DV	0.923	0.923	0.923	0.923
Policy FE	Yes	Yes	Yes	Yes
Year FE	No	Yes	Yes	Yes
Round FE	No	No	Yes	Yes
Politician control	No	No	No	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. We estimate the following model: $Similarity_{ip} = \beta \cdot Incentive_{it} + \sigma_p + \theta_{t(p)} + X_i\Gamma + \varepsilon_{ipt}$. Career incentives are measured as the ex-ante probability of promotion projected by the start age of tenure and hierarchical level (mean=0.481, s.d.=0.075). For the outcome variable (y_{ip}), we conduct Latent Semantic Analysis, an approach from Natural Language Processing, to measure the text similarity of government documents. The similarity index, taking maximum over all similarity pairs between a document issued by prefecture i, 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. The mean of the outcome variable is 0.922, with a standard deviation of 0.081. 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 (X_i) include his or her level of education and previous central experience. Standard errors are clustered at the policy level.

Table A.14: Land revenue windfall and experimentation roll-out - first stage

	Land revenue		
	(1)	(2)	(3)
Suitability$\times$ interest rate	3.062***	3.176***	3.181***
	(0.118)	(0.126)	(0.127)
# of obs.	66,128	66,128	66,128
# of clusters	1644	1644	1644
Mean of DV	5.271	5.271	5.271
Ministry FE	No	No	Yes
Year FE	Yes	Yes	Yes
County FE	No	Yes	Yes

Note: This table presents the first stage estimates for the two-stage-least-square regression in Table 3, Panel A. Specifically, we estimate the following model using county-policy-year level data: $Land\_revenue_{ipt} = \alpha \cdot Suitability_i \times Interest_t + X'_{it}\beta + \delta_i + \gamma_t + \delta_{m(p)} + \epsilon_{impt}$. The outcome of interest is the average land revenue collected, across the entire experimentation period, in logarithm terms. The instrument variable is an interaction between % of suitable land for conversion (angle of slope $0 - 15^{\circ}$) and national interests rate (mean=0.150, s.d.=0.325). Following Chen and Kung (2016), we include politician level controls (age, education, past experience in the prefectural government, previous positions as Youth League party leaders, and hometown connection with the prefectural leaders). Standard errors are clustered at the county level.

Table A.15: Naive evaluation of policy experimentation: county-year level analyses

	National roll-out		
	(1)	(2)	(3)
Panel A: Land revenue windfall			
Land revenue (instrumented)	0.009**	0.021***	0.011***
	(0.003)	(0.004)	(0.003)
First stage F stats	229.80	207.09	154.15
# of obs.	9,049	9,049	9,049
# of clusters	1514	1470	1470
Mean of DV	0.358	0.358	0.358
Year FE	No	No	Yes
County FE	No	Yes	Yes
Controls	Yes	Yes	Yes
Panel B: Political rotation			
Rotation	-0.005	-0.007	-0.004
	(0.009)	(0.009)	(0.008)
Rotation × change in career incentive	0.309**	0.237*	0.284**
	(0.129)	(0.127)	(0.116)
# of obs.	5806	5802	5802
# of clusters	294	290	290
Mean of DV	0.146	0.146	0.146
Year FE	No	No	Yes
County FE	No	Yes	Yes
Controls	Yes	Yes	Yes

Note: In panel A, we estimate the following 2 stage least square model.

$$Land\_revenue_{it} = \alpha \cdot Suitability_i \times Interest_t + X'_{it}\beta + \delta_i + \gamma_t + \epsilon_{it}$$

$$y_{it} = \mu \cdot \widehat{Land\_revenue}_{it} + X'_{it}\Gamma + \psi_i + \nu_t + \varepsilon_{it}$$

, where y_{it} is the percentage of policies being rolled out in county i in year t. 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 (mean=3.96, s.d.=3.60). 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. The standard errors are clustered at the county level. Panel B is an analysis focusing on political rotations that happened *after* the selection of experimentation sites. At the prefecture-by-year level, we calculate the difference in career incentives between the leaving prefectural official and his immediate successor. We then estimate the following model: $y_{it} = \beta_1 \cdot Rotation_{it} + \beta_2 \cdot Rotation_{it} \times \Delta incentive_{it} + \delta_i + \gamma_t + \varepsilon_{it}$. *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. 23.9% of the participating prefectures went through politician rotation *during* the experimentation period. Career incentives are measured as the ex-ante probability of promotion projected by the start age of tenure and hierarchical level. The standard errors are clustered at the province level.

Table A.16: Naive evaluation of policy experimentation

	National roll-out		
	(1)	(2)	(3)
Panel A: Baseline			
Land revenue (instrumented)	0.008***	0.006***	0.009***
	(0.002)	(0.001)	(0.001)
Panel B: Weighted by experimentation sites			
Land revenue (instrumented)	0.012***	0.020***	0.018***
	(0.004)	(0.005)	(0.005)
Panel C: Small scale experiments			
Land revenue (instrumented)	0.015***	0.008***	0.009***
	(0.002)	(0.002)	(0.002)
# of obs.	66,128	66,128	66,128
# of clusters	1,644	1,644	1,644
Experiment Year FE	Yes	Yes	Yes
County FE	No	Yes	Yes
Ministry FE	No	No	Yes

Note: Standard errors clustered at the county level are reported below the parentheses. Panel A reproduces our full-sample estimates from Table 3, Specifically, we estimate the following econometric model:

$$Land\_revenue_{ipt} = \alpha \cdot Suitability_i \times Interest_t + X'_{it}\beta + \delta_i + \gamma_t + \delta_{m(p)} + \epsilon_{impt}$$

$$y_p = \mu \cdot \widehat{Land\_revenue}_{ipt} + X'_{it}\Gamma + \psi_i + \nu_t + \delta_{m(p)} + \varepsilon_{ipmt}$$

Panel B runs the same regression, but applied weights using the inverse of experimentation sites. Smaller experiments get larger weights. Panel C runs the same regression on smaller-scale experiments with less-than-average number of participating localities (N=45,165).

Table A.17: Land revenue windfall and experimentation's national roll-out: placebo exercises

	National roll-out		
	(1)	(2)	(3)
Land revenue (instrumented by baseline IV)	0.014***		
	(0.002)		
Land revenue (instrumented by placebo IV)		0.281	
		(0.303)	
Land revenue$_{t+5}$ (instrumented by IV$_{t+5}$)			-0.002
			(0.004)
First stage F stat	474.60	1.12	379.80
# of obs.	16782	7970	12480
# of clusters	1642	1642	1642
Mean of DV	0.358	0.358	0.358
County FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes

Note: This table shows placebo tests for land revenue IVs. We estimate the following econometric model:

$$Land\_revenue_{ipt} = \alpha \cdot Placebo\_IV_{it} + X'_{it}\beta + \delta_i + \gamma_t + \delta_{m(p)} + \epsilon_{ipmt}$$

$$y_p = \mu \cdot \widehat{Land\_revenue}_{ipt} + X'_{it}\Gamma + \psi_i + \nu_t + \delta_m + \varepsilon_{ipmt}$$

Standard errors clustered at the county level are reported below the estimates. In column 1, we report our baseline results in the cross-sectional regression: instrumenting land revenue with the shift-share IV (% of suitable land for conversion (angle of slope $0-15°$) interacted with national interests rate, mean=0.150, s.d.=0.325). In column 2, we report the second stage estimates with a placebo IV where, instead of the % of suitable land for conversion (angle of slope $0-15°$), we use the % of steeper agrarian lands within a county (angle of slope $15-30°$). In column 3, we report the second stage estimates with the instrumented 5-years-lagged land revenue.

Table A.18: Scale of the experiment and rotation effect

	Rollout	
	(1)	(2)
Rotation	0.086**	0.032**
	(0.040)	(0.013)
Positive rotation × Δ incentive	0.524**	0.004
	(0.207)	(0.073)
Positive rotation × Δ incentive × small-scale exp		0.487***
		(0.131)
Negative rotation × Δ incentive	-0.586***	-0.100
	(0.201)	(0.099)
Negative rotation × Δ incentive × small-scale exp		-0.339*
		(0.190)
Small-scale experiments		0.376***
		(0.023)
# of obs.	3,890	3,890
# of clusters	27	27
Mean of DV	0.321	0.321
Weighted	Yes	No
Year FE	Yes	Yes
Province FE	Yes	Yes
Ministry FE	Yes	Yes

Notes: Standard errors clustered at the province level are reported in the parentheses. Our empirical setup follows Table 3, Panel B. Specifically, we estimate the following model: $y_p = \alpha \cdot Turnover_{ip} + \beta_1 \cdot Turnover_{ip} \times IncreaseIncentive_{ip} + \beta_2 \cdot Turnover_{ip} \times DecreaseIncentive_{ip} + \gamma_t + \delta_{m(p)} + \theta_n + \varepsilon_{ipmnt}$. In column 1, we weigh the policy × experimental-site-level regression with the inverse number of experimentation site. In column 2, we include an indicator of less-than-average experimentation sites as an interaction term. At policy-by-prefecture level, 53.9% of participating localities went through rotation. An average positive rotation is accompanied with an incentive increase of 0.079 (s.d.=0.076). An average negative rotation is accompanied with an incentive drop of 0.055 (s.d.=0.061).

Table A.19: Political rotation and experimentation's national roll-out

	National roll-out		
	(1)	(2)	(3)
Panel A: Alternative measure of career incentive			
Rotation	-0.002	0.010	0.016
	(0.027)	(0.019)	(0.018)
Rotation × increase in career incentive	0.146***	0.106***	0.061*
	(0.021)	(0.033)	(0.033)
Rotation × drop in career incentive	-0.115***	-0.082*	-0.088*
	(0.033)	(0.044)	(0.049)
Panel B: Low-stake policies			
Rotation	-0.024	-0.023	-0.002
	(0.030)	(0.016)	(0.018)
Positive rotation ×Δ Incentive	0.641***	0.469***	0.400**
	(0.161)	(0.122)	(0.159)
Negative rotation ×Δ Incentive	-0.439***	-0.338***	-0.248*
	(0.160)	(0.112)	(0.135)
Panel C: Last minute rotation			
Rotation	-0.046*	-0.037**	-0.027*
	(0.025)	(0.018)	(0.016)
Rotation × increase in incentive	0.790***	0.686***	0.588***
	(0.204)	(0.161)	(0.154)
Rotation × decrease in incentive	-0.357*	-0.334**	-0.282**
	(0.187)	(0.138)	(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

Table A.19: Political rotation and experimentation's national roll-out

	National roll-out		
	(1)	(2)	(3)
Panel D.1: Pre-experiment rotation			
Pre-exp rotation	-0.009	-0.017	-0.022
	(0.023)	(0.015)	(0.015)
Pre-exp rotation × increase in career incentive	0.376	0.227	0.229
	(0.235)	(0.183)	(0.186)
Pre-exp rotation × drop in career incentive	-0.204	-0.127	-0.100
	(0.140)	(0.096)	(0.102)
Panel D.2: Post-experiment rotation beyond 5 years			
Rotation at $T+5$	0.049*	-0.001	0.004
	(0.027)	(0.014)	(0.014)
Rotation at $T+5$ × increase in career incentive	0.057	0.211	0.185
	(0.173)	(0.189)	(0.214)
Rotation at $T+5$ × decrease in career incentive	-0.216	-0.186	-0.123
	(0.254)	(0.139)	(0.160)
# of obs.	4,670	4,659	3,890
# of clusters	27	27	27
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 below the estimates. In general, this table follows the baseline specifications in Table 3, Panel B. Specifically, we estimate the following model: $y_p = \alpha \cdot Turnover_{ip} + \beta_1 \cdot Turnover_{ip} \times IncreaseIncentive_{ip} + \beta_2 \cdot Turnover_{ip} \times DecreaseIncentive_{ip} + \gamma_t + \delta_{m(p)} + \theta_n + \varepsilon_{ipmnt}$. In panel A, we adopt an alternative definition to compute career incentives. We define a rotation with increase in career incentive to occur when the last politician who is more than 58 years old is replaced by a young successor starting before 57, and vice versa. Consistent with our sample definition in Table A.6, we exclude politicians younger than age 50 to estimate a 'local effect'. In panel B, we focus on the subsample of policy experiments that are relatively low-stake, defined as not appearing in the Five-Year-Plans prior to the experiments. In panel C, we focus on rotations that happen in the last year of the experiment. In panel D.1, we instead consider political rotations before experimentation starts, while in panel D.2, we consider rotations that happen 5 years after the experimentation starts.

Table A.20: Experimentation effects and national roll-out

	National roll-out		
	(1)	(2)	(3)
Panel A: Pre vs. post comparison			
Experiment effect	0.041**	0.041**	0.070**
	(0.019)	(0.019)	(0.033)
Panel B: Controlling for pre-trend			
Experiment effect	0.013	0.002	0.015
	(0.026)	(0.030)	(0.032)
Panel C: Synthetic control			
Experiment effect	0.029	0.034	0.012
	(0.042)	(0.046)	(0.086)
# of obs.	355	355	355
# of clusters	44	44	44
Mean of DV	0.390	0.390	0.390
Evaluation year FE	Yes	Yes	Yes
Ministry FE	No	Yes	No
Minister FE	No	No	Yes

Note: This table examines the association between experimentation effects, estimated in a variety of ways, and whether the corresponding experiment leads to the policy's national roll-out. The analysis is conducted at policy level. Specifically, we estimate the model $Rollout_i = \beta \cdot \widehat{ATE_{exp}}_i + \gamma_m + \theta_t + \varepsilon_{imt}$ In order to make magnitudes comparable across panels, all independent variables are standardized to mean zero and unit variance. In Panel A, we regress policy roll-out on the simple pre-vs-post estimates based on experimentation sites' GDP per capita. In panel B, we in addition controls for experimentation sites' county specific pre-trend. In Panel C, we follow Xu (2017) and conduct a generalized synthetic control analysis to estimate the experimentation effect, matching 3-year trend in local socioeconomic conditions prior to the experimentation. All columns control for year of evaluation fix effects (θ_t); column 2 in addition control for ministry fixed effect; and column 3 controls for minister fixed effects instead. Standard errors are clustered at the ministry level. Sample: all economic policy experiments.

Table A.21: Winsorized experiment effect and policy rollout

	Rollout		
	(1)	(2)	(3)
Average experiment effect	0.0703**	0.0547	0.0709*
	(0.0338)	(0.0378)	(0.0424)
# of obs.	293	293	293
Mean of DV	0.387	0.387	0.387
Minister FE	Yes	Yes	Yes
Evaluation year FE	Yes	Yes	Yes
Winsorize	None	bottom 2.5%	top 2.5%

Notes: In this table, we follow our main specification to estimate the model, but with a winsorized sample. Specifically, the empirical setup is $Rollout_i = \beta \cdot \widehat{ATE_{exp_i}} + \gamma_m + \theta_t + \varepsilon_{imt}$. Robust standard errors are reported below the estimates. Column 1 follows the exact same specification as Figure 4, panel A in our paper, whereas in column 2 and 3 we winsorize extreme values on each end respectively. The independent variables are standardized to mean zero and unit variance. Sample: all economic policy experiments.

Table A.22: Similarity with experimentation sites and effects of policy roll-out: robustness checks

	GDP per capita growth		
	(1)	(2)	(3)
Panel A: GDP per capita			
M-distance between local development	-0.004***	-0.004***	-0.003***
	(0.001)	(0.001)	(0.001)
Panel B: GDP per capita + fiscal income + fiscal expenditure			
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***	-0.003***	-0.003***
	(0.001)	(0.001)	(0.001)
# of obs.	94,635	94,635	94,635
# of clusters	2064	2064	2064
Mean of DV	0.102	0.102	0.102
Policy FE	Yes	No	Yes
County FE	No	Yes	Yes

Note: This table presents results repeating the specification in Table 4, but with alternative formulations of the Mahalanobis distance. Specifically, we estimate the following model: $Growth_{cpt} = \alpha \cdot M_{cp} + \gamma_c + \sigma_t + \eta_p + \epsilon_{cpt}$. We compute the M-distance using a variety of measures. Panel A focuses on similarity in GDP per capita; Panel B focuses on similarity in GDP per capita, local fiscal income and local fiscal expenditure; and Panel C focuses on GDP per capita, local fiscal income, and local population size. All independent variables are standardized to mean zero and unit variance. Standard errors are clustered at the county level.

Table A.23: Political patronage and engagement in experimentation

	Engaged in experimentation		
	(1)	(2)	(3)
Panel A: All experiments			
Connected to minister	0.088**	0.062*	0.063*
	(0.035)	(0.036)	(0.037)
Panel B Experiments with top-down assignments			
Connected to minister	0.073**	0.056*	0.058*
	(0.029)	(0.031)	(0.031)
Panel C: Experiments with voluntary sign-ups			
Connected to minister	0.015	0.006	0.006
	(0.018)	(0.016)	(0.016)
# of obs.	42884	42884	42884
# of clusters	31	31	31
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: This table reports the estimates of the following econometric model using ministry-province-year level data: $y_{mpt} = \alpha \cdot Connection_{mpt} + \delta_{mp} + \theta_t + \varepsilon_{mpt}$, 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; θ_t is year fixed effects; and δ_{mp} stands for province-by-ministry fixed effects. "Connected to minister" is an indicator variable showing whether the current minister had worked in the province where the experiment took place (mean=0.019, s.d.=0.136). In column 3, we in addition control for the provinces' value added of first and second industry, fiscal expenditure and income of local governments. The standard errors are clustered at the province level.

Table A.24: Concerns for political stability and selection of experimentation sites

	Engaged in experimentation		
	(1)	(2)	(3)
# of protests in previous year	-0.004***	-0.003***	-0.003***
	(0.001)	(0.001)	(0.0002)
# of obs.	1730	1730	940
# of clusters	190	190	190
Mean of DV	1.278	1.278	2.117
Prefecture FE	Yes	Yes	Yes
Year FE	No	Yes	Yes
Pre-period controls	No	No	Yes

Note: This table reports associations between localities' engagement in policy experimentation and occurrence of protests in the locality during previous year (mean=4.43, s.d.=42.65). Specifically, we estimate: $y_{pt} = \alpha \cdot protest_{p,t-1} + \delta_p + \theta_t + \varepsilon_{pt}$, where y_{pt} is a dummy variable indicating whether prefecture p engages in policy experimentation in year t (measured with "event" counts in the GDELT database (Beraja et al. 2023)); $protest_{p,t-1}$ is the number of protests occurred in prefecture p in year $t-1$; δ_p is prefecture fixed effects; and θ_t is year fixed effects. A full set of prefecture fixed effects are controlled across all columns; columns 2 and 3 in addition control for year fixed effects; and column 3 in addition controls for localities' GDP per capita in the previous year. The standard errors are clustered at the prefecture level.

Table A.25: Concerns for political stability and policies' national roll-out

	National roll-out		
	(1)	(2)	(3)
Panel A: OLS			
Protest	-0.007*	-0.007*	-0.007*
	(0.004)	(0.004)	(0.004)
Panel B: Instrumented protest			
Protest	-0.020***	-0.020***	-0.018***
	(0.005)	(0.005)	(0.005)
First stage F stats	63.18	62.87	65.01
# of obs.	1,122	1,122	1,122
# of clusters	214	214	214
Mean of DV	0.747	0.747	0.747
Controlling for ATE	Yes	Yes	Yes
Controlling for economic conditions	No	No	Yes
Prefecture FE	Yes	Yes	Yes
Evaluation year FE	Yes	Yes	Yes
Ministry FE	Yes	No	No
Minister FE	No	Yes	Yes

Note: Standard errors are clustered at the prefecture level. This table examines whether occurrence of protests during experimentation is associated with the experimental policies' roll-out to the entire nation. The analysis is conducted at policy-prefecture level. "Protest" is the count of protests occurred during the experimentation period, aggregated via GDELT data from 2014 to 2020 (mean=3.28, s.d.=6.04). We run a policy-prefecture level regression: $y_i = \beta Protest_{pt} + \delta_p + \gamma_t + \epsilon_{ipt}$, where y_i is an indicator of whether the policy is rolled out to the entire nation, $Protest_{pt}$ is the number of protests in prefecture p in its beginning year t. We control for prefecture fixed effects δ_p, and year fixed effects γ_t. Results are winsorized at 95 percentile to get rid of extreme values, but will be qualitatively robust without winsorization. "ATE" is the average treatment effect, calculated as the pre-vs-post comparison in local GDP per capita. In panel A, we presented OLS estimates, and in panel B, we instrument protest with rain, gust, winds and their interaction with protests elsewhere, following the parsimonious IV specification of Beraja, Yang, and Yuchtman (2023). In column 1-3 we control for different sets of fixed effects, and in column 4 we control for local GDP, local city population and local fiscal revenue.

Table A.26: Ministries completed 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.27: Political career incentives and engagement in experimentation

	# of experiments engaged		
	(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)
# of obs.	7630	7630	7630
# of clusters	334	334	334
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: This table examines the relationship between politicians career incentives and localities' engagement in policy experiments. We estimate the following econometric model: $y_{pt} = \alpha \cdot Incentive_{pt} + X'_{pt} \cdot \beta + \delta_p + \theta_t + \varepsilon_{pt}$, where y_{pt} is the number of experiments occurring in prefecture p and year t. Panel A reports the estimated effect of career incentive intensity on all types of experimentation. Panel B distinguishes between experiments issued by a M-form ministry (where city leaders have direct control on the logistics of the policy) and U-form ministry (where the central government takes direct orders on local branches). 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. Control variables at the politician level include 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. Average career incentive is 0.474, and the standard deviation is 0.082. Standard errors are clustered at the prefecture level.

Table A.28: Political career incentives and engagement in experimentation: placebo exercise

	Engaged 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
# of clusters	333	333	333
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: This table presents a placebo test to the last table, exploring the relationship between the previous politician's career incentive and localities' engagement in policy experiments. We estimate the model $y_{pt} = \alpha \cdot Placebo\_incentive_{pt} + X'_{pt} \cdot \beta + \delta_p + \theta_t + \varepsilon_{pt}$ We construct the same career incentive indices (mean=0.474, s.d.=0.082.), but we replace in-office prefecture leaders' career incentives with those of their immediate predecessors. 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. Control variables at the politician level include career incentive of the previous city leader to address the concern where the engagement is just a continuation of previous progress. Standard errors are clustered at the prefecture level.

Table A.29: 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)
# of obs.	2,337	2,337	2,337	2,337
Mean of DV	0.637	0.637	0.637	0.637
Controls	No	Yes	No	Yes

Note: In this table, we show how we constructed our career incentive variable following Wang, Zhang, and Zhou (2020). Specifically, we estimate a probit regression $\hat{y}_{pt} = \Phi^{-1}\left\{\hat{\alpha} \cdot startage_{pt} + \hat{\beta} \cdot level_{pt} + \hat{\gamma} \cdot startage_{pt} \times level_{pt}\right\}$.. Average start age of prefecture politicians is 48.85 (s.d.=4.19). Most politicians are zhengtin level, but 7.6% are higher up in political hierarchy (fubu). Columns 1 and 2 report OLS estimates, and the next two columns report estimates from a probit regression. 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. Robust standard errors are reported below the estimates.

Table A.30: Political incentives and engagement in experimentation

	Engaged in experimentation			
	(1)	(2)	(3)	(4)
GDP per capita	0.022***	0.010***	0.045***	0.026***
	(0.001)	(0.001)	(0.003)	(0.002)
# of obs.	68,335	70,237	68,335	70,237
# of clusters	250	250	250	250
Mean of DV	0.023	0.023	0.023	0.023
Controls for political incentives	No	Yes	No	Yes
Policy FE	No	No	Yes	Yes

Note: This table examines how much of experimentation sites' positive selection may be attributed to misaligned political incentives, where we regress localities' engagement in experimentation on their pre-experimentation GDP per capita, in logarithm (mean=0.819, s.d.=0.918). Specifically, we estimate the following econometric model: $y_{ip} = \beta \cdot GDP_{ip} + \gamma_p + (X_{ip}\Gamma) + \varepsilon_{ip}$. Columns 1 and 3 do not control for local political incentives (X_{ip}), and columns 2 and 4 include controls for local political incentives (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 D.2). Columns 3 and 4, in addition, control for a full set of policy fixed effects. This analysis is conducted in a a subsample of experiments targeting prefectural cities only, since political incentives are measured at the prefecture level. Standard errors are clustered at the policy level.

Table A.31: Total fiscal expenditure during policy experimentation

	Total fiscal expenditure		
	(1)	(2)	(3)
# of experiments	3,742.465***	3,038.238***	1,215.933***
	(188.987)	(65.095)	(78.897)
# of obs.	25,196	25,196	25,196
# of clusters	1,973	1,973	1,973
Mean of DV	15098	15098	15098
County FE	No	Yes	Yes
Year FE	Yes	No	Yes

Note: This table is a simplified version of our analysis on fiscal allocation, where we estimate the following econometric model: $y_{it} = \beta \cdot Exp_{it} + \gamma_i + \theta_t + \varepsilon_{it}$. Instead of focusing on domain-specific fiscal expenditure, we investigate whether participating in experimentation is associated with localities' *total* fiscal expenditure, in 10,000 Yuan (s.d.=13832). On average, localities participate in 1.26 policy experiments each year, with a standard deviation of 1.44. Standard errors are clustered at the county level.

References

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

Beraja, Martin, David Y Yang, and Noam Yuchtman. 2023. "Data-intensive innovation and the State: evidence from AI firms in China." *The Review of Economic Studies* 90 (4): 1701–1723.

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.*

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

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.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics* 225 (2): 175–199.

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

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

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