Policy Experimentation in China: The Political Economy of Policy Learning

Shaoda Wang

University of Chicago, National Bureau of Economic Research, and Bureau for Research and Economic Analysis of Development

David Y. Yang

Harvard University, National Bureau of Economic Research, Bureau for Research and Economic Analysis of Development, Abdul Latif Jameel Poverty Action Lab, and Canadian Institute for Advanced Research

Governments use policy experiments to facilitate learning, but the nature and effects of these experiments remain unclear. We analyze China's policy experimentation since 1980—among the most systematic in history—and document three facts. First, most experiments exhibit positive sample selection. Second, local politicians exert excessive efforts during experiments that are not replicable during policies' national rollout. Third, the central government is not fully sophisticated when interpreting experimentation outcomes. These facts suggest that policy learning may be biased and national policies may be distorted. Thus, while China's institutions enable experimentation at an unparalleled scale, the complex political environments can also limit effective policy learning.

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, Anran Tan, and in particular Kaicheng Luo provided outstanding research assistance. This paper was edited by John A. List.

Electronically published May 7, 2025

Journal of Political Economy, volume 133, number 7, July 2025.

https://doi.org/10.1086/734873

I. Introduction

Determining which policies to implement and how to implement them is an essential task for any government (e.g., Hayek 1978; North 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 multifaceted (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 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 subregions of a country. Few, however, can compare to the systematic policy experimentation in China in terms of its breadth, depth, and duration. Since the 1980s, the Chinese government has been 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 the 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 rollout 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 preexperimentation 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 nonexperimentation 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 rollout—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 rollout. 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 (ATE; when policies are implemented to the average locality with average local politician incentives),[1] the presence of positive sample selection

[1] In sec. VIII, 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

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 sec. IV 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 rollout. 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 persuasion across decision-makers are intertwined; and Callander and Harstad (2015) investigate how decentralized jurisdictions strategically engage in policy experimentation and how a central

that considers the central government's demand for political stability during and after policy experimentation.

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, 2021; Davis et al. 2017; Al-Ubaydli, List, and Suskind 2019; List 2022). Consistent with the theoretical framework proposed by Al-Ubaydli, List, and Suskind (2019), we find that both nonrepresentative experimental samples and nonrepresentative 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 United States, as well as findings by DellaVigna and Linos (2020) on trials conducted by the Nudge Units in the United Kingdom 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 selection bias in conditional cash transfer and microcredit experiments initiated by the Jameel Poverty Action Lab or Innovations for Poverty Action (Gechter

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 II provides institutional background on China's policy experimentation. Section III describes the data sources, the process of constructing the database on policy experimentation, and a number of key characteristics on policy experimentation. Section IV 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 (sec. V), strategic efforts by local politicians during the experiments (sec. VI), and nonsophisticated interpretation of experimentation outcomes (sec. VII). These can be interpreted without taking a stance on the central government's objectives for experimentation. Section VIII discusses the implications for learning from experimentation and national policy outcomes, under specific assumptions about the government's objectives. Finally, section IX concludes.

II. 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, 3). The central government plays a key role in initiating and coordinating policy experimentation. While China's economic reforms are often accompanied by decentralization, powerful political centralization remains a key characteristic

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).

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, 3). 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 and Treisman 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 preconceived, 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, 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: "The 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" (11).

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 cosigned by coordinating ministries or the State Council if interministerial 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, 407).[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 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

[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.

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 section A.1 (the appendix is available online) 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 (2002–15), carbon emission trading (2011–21), separation of permits and licenses (2015–18), and introduction of agriculture catastrophe insurance (2017–21). We describe these experiments in greater detail in section III.C.

III. 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 III.A. 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 section B. In section III.B, we present 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 III.C.

A. Government Documents on Policy Experimentation

Our main data are based on the comprehensive collection of policy documents issued by the Chinese central and local governments since 1949, compiled by PKU Law (https://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 keywords “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 documents mark the official initiation of particular policy experiments, their key milestones (e.g., when a major expansion of experimentation is planned), and decisions to roll out the policies to the entire country if the experiment is successful. Local government documents are issued by each locality participating in the experiments, specifying details on local 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 rollout, and 39 are policies with specific regional targets.[6] For the baseline analysis in section V, 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 secs. VI–VIII).

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 include 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 PKU Law has extensive and comprehensive coverage (see table A.2; tables A.1–A.31 are available online). When we manually examine the limited documents that are published on the ministries' websites but not included in the PKU Law database, we find that they are secondary documents and do not contain additional critical information.

[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 antipoverty 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.

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 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—for example, 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.

B. 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 multiyear rollout 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 fewer 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, multidivisional 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

[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.

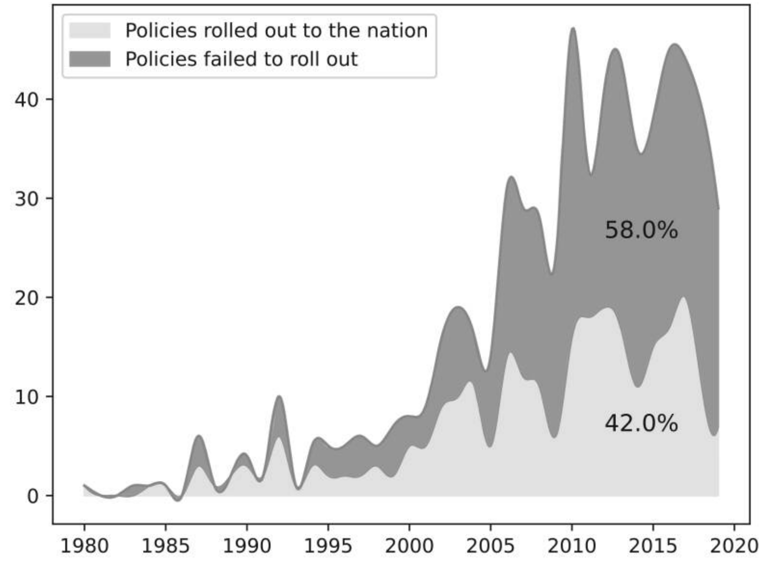
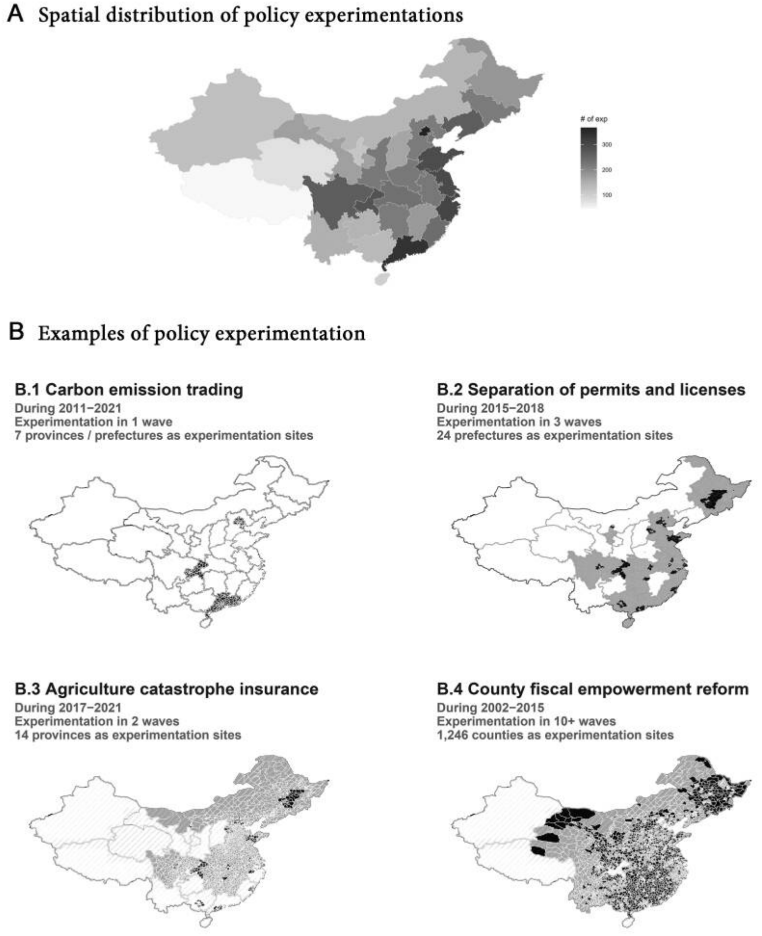


FIG. 1.—Number of policy experiments initiated over time. The share of successful experiments that eventually rolled out to the entire country is indicated by the area shaded in lighter gray; the share of unsuccessful policies that failed to roll out to the entire country is indicated by the area shaded in darker gray.

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 section C presents results using an event study design.

Experimentation sites.—We extract the experimentation sites of each policy experiment.[8] Figure 2*A* plots the distribution of experimentation sites across China, aggregated at the province level (for county-level distribution, see fig. A.1; figs. A.1–A.35 are available online). Panel A of table 1 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.

[8] Many policy experiments have more than one wave of rollout, and we identify 1,374 distinct rounds of rollout 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 rollout and lasts for 2.25 years, until either the rollout stops or the experiment becomes a national policy. We leave it to future work to systematically study the dynamic experimentation implementation.



A Spatial distribution of policy experimentations

B Examples of policy experimentation

B.1 Carbon emission trading
During 2011–2021
Experimentation in 1 wave
7 provinces / prefectures as experimentation sites

B.2 Separation of permits and licenses
During 2015–2018
Experimentation in 3 waves
24 prefectures as experimentation sites

B.3 Agriculture catastrophe insurance
During 2017–2021
Experimentation in 2 waves
14 provinces as experimentation sites

B.4 County fiscal empowerment reform
During 2002–2015
Experimentation in 10+ waves
1,246 counties as experimentation sites

FIG. 2.—*A*, Total number of policy experiments that each province has been involved in between 1980 and 2020 (including experiments at prefectural and county levels). *B.1*, *B.2*, Two policies that eventually rolled out to the entire country. The regions shaded in black indicate parts of the country that eventually received the policies when they rolled out. *B.3*, *B.4*, Two policies that did not eventually roll out, indicated by a striped background. The experimentation sites are marked in black, and the corresponding provinces are marked in gray.

About 42.6% of the experiments allow (at least partially) for voluntary participation of the local governments (see fig. A.2).

National policy rollout.—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

TABLE 1
Summary Statistics of Policy Experimentation

	Number of Experiments (1)	Number of Rounds (2)	Number of Sites (3)	% Rollout (4)	Average *t*-Statistics (5)	% Representative (6)
	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
National × completed	509	2.1	18.8	50.88	5.48	39.4
National × ongoing	104	2.1	23.9	8.65	3.58	56.3
Subnational	39	2.0	8.7	15.38	2.22	60.0
Subnational × completed	35	2.0	9.2	17.14	2.16	59.3
Subnational × ongoing	5	2.0	11.4	0.00	2.72	50.0
	B. By Policy Domain					
Resource, energy, and 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 and fiscal policy	41	2.2	10.2	53.66	5.38	38.2
Population and health	38	2.3	21.2	47.37	4.57	36.1
Commerce and trade	36	2.1	17.0	41.67	6.34	23.1
Industry and 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 and 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
	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 subnational ones (39). In panels B and C, we focus only on those national experiments.

became national policies, while 58.0% failed (see fig. 1; share of successful and failed experiments indicated by darker and lighter gray shades, respectively). The share of policy experimentation leading to national policy rollout remains remarkably stable over time (see fig. A.3). The patterns concerning policies' national rollout 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 fig. 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 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. Panel B of table 1 presents the number of policy experiments initiated by different ministries and commissions, grouped by policy domains and broad functions for which they are responsible. 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 consider a policy on trial to be more certain based on several criteria: (i) whether the central government has laid out a detailed national rollout timeline before the experiment starts,[9] (ii) whether experimentation details were already drafted out by the central government at the beginning of the experiment, or (iii) whether 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 before the beginning of the experiment; we calculate the average textual similarity across these papers (using term frequency–inverse document frequency).

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) whether multiple ministries and commissions are involved in the experiment,[10] (ii) whether the government documents describing the experiments are long and/or contain multiple documents,

[9] Of all the experiments, 30.8% 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 fig. A.6).

[10] Of all the policy experiments, 23.8% involve more than two ministries and commissions; we label these as complex experiments (see fig. A.7)

(iii) whether the experiment duration is long, or (iv) whether 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 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]

C. 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 app. sec. A.2 for additional details of those examples). In addition, in table A.1 we present examples of policy experiments across a variety of policy domains.

Figure 2*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. Figure 2*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.

Figure 2*B.3* and 2*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 fig. 2*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 figure 2*B.4*, the experimentation on county fiscal empowerment took place

[11] Text similarity is calculated by a pretrained model from PaddlePaddle, and we provide more details in app. sec. G.

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.

IV. 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 represents the policy outcome of interest,[12] p corresponds to the specific policy of interest, $\mathbb{I}_t$ represents the preexperimentation 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 decompose the observed experimentation outcomes as follows:

$$\hat{Y}(p, \mathbb{I}_t, \mathbb{E}_t) = \underbrace{\bar{Y}(p, \bar{i}_t, \bar{e}_t)}_{\text{ATE}} + \underbrace{F_{i,p}(\mathbb{I}_t - \bar{i}_t)}_{\text{site selection}} + \underbrace{F_{e,p}(\mathbb{E}_t - \bar{e}_t)}_{\text{experimental situation}} + G_{i,e}(\mathbb{I}_t, \mathbb{E}_t), \quad (1)$$

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 ATE may be a parameter of interest to the policymaker because it indicates the expected outcome of the policy on trial when it is 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 V documents nonrepresentative 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 $\bar{Y}$

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

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 VI aims to document the presence of nonrepresentative experimental situations—that is, $\mathbb{E}_t - \bar{e}_t \neq 0$.

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

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

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 nonrepresentative sample selection and nonrepresentative 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 nonrepresentative 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 (e.g., an objective function that gives more weight to the economic performance of certain localities in the country).

[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.

The term δ_2 captures the possibility that the presence of nonrepresentative sample selection and nonrepresentative 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 VII aims to document that the central government is indeed not fully sophisticated when interpreting experimentation 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 VIII.

V. 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 IV, we test whether $\mathbb{I}_t - \bar{i}_t = 0$.

A. *Procedure to Test for Representativeness*

For each policy experiment, we compare preexperimentation 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 preexperimentation levels of local fiscal expenditure are indistinguishable among the experimentation sites and nonexperimentation sites. This amounts to 652 separate t-tests, one for each policy experiment.[14] In section V.B, 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

[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.

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

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

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 nonexperimenting counties or prefectures within the same province.[16]

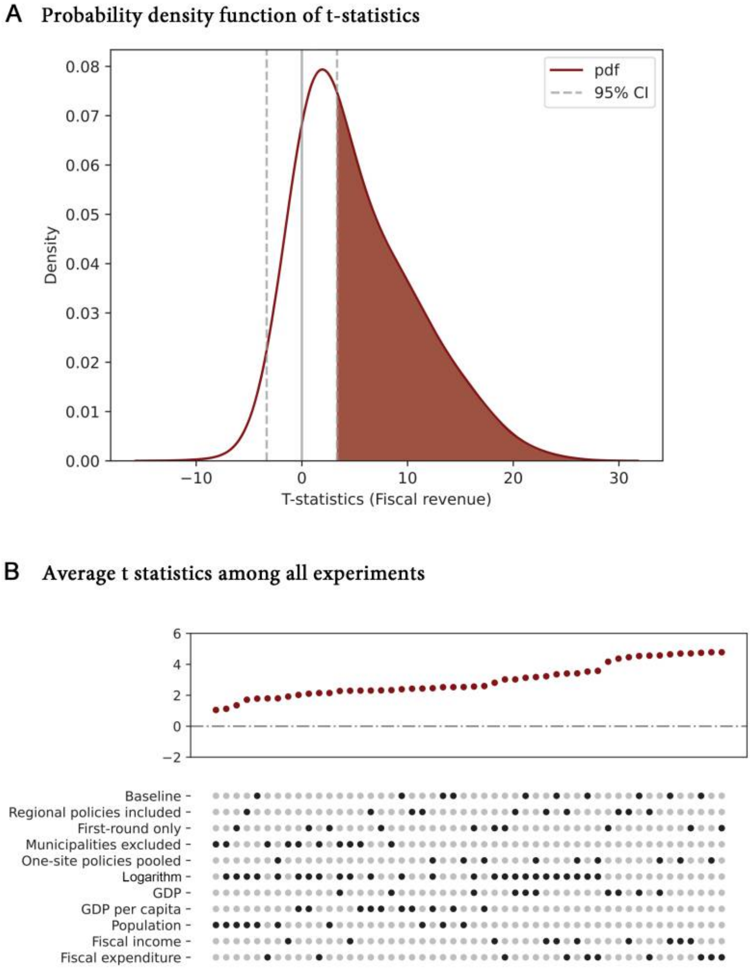
Second, approximately one-quarter of the experiments involve only one experimentation site. We cannot conduct standard statistical tests for these single-site experiments. Instead, we pool each single-site experiment with four other randomly selected single-site experiments and conduct the representativeness test on the pooled sample, where the nonexperimentation 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.

B. *Most Experimentation Sites Are Positively Selected*

In figure 3*A*, we plot the distribution of the baseline t-statistics comparing preexperimentation local fiscal revenue between the experimentation and nonexperimentation sites. We mark the thresholds of t-statistics where one would reject the null hypothesis of representative site selection

[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. Specifically, $\nu_i = (s_{i,1}^2/n_{i,1} + s_{i,2}^2/n_{i,2})^2/((s_{i,1}^2/n_{i,1})^2/(n_{i,1} - 1) + (s_{i,2}^2/n_{i,2})^2/(n_{i,2} - 1))$.

[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.


A Probability density function of t-statistics

B Average t statistics among all experiments

FIG. 3.—*A*, 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 preexperimentation 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. *B*, Summary of *t*-test results for other specifications. We present those results as a specification curve à la Simonsohn, Simmons, and Nelson (2020), with filled circles indicating the combination of plausible specifications.

at the 95% confidence interval.[17] Panel A of table 1 reports the corresponding test statistics (adjusting for the degrees of freedom for each

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

test) and the share of policy experiments for which we can reject the null hypotheses at the 5% significance level (in cols. 5 and 6).

We find that the average of the *t*-statistics comparing experimentation and nonexperimentation sites is 5.17, across all experiments that are intended for national rollout. 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 are intended for national rollout, we (i) include experiments on policies that target specific regions and adjust the nonexperimentation sites according to the regional scope; (ii) focus on only 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 in the 5 years before the experiment, (iv) local population, and (v) local fiscal expenditure. In figure 3*B*, we plot the average *t*-statistics comparing experimentation and nonexperimentation 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 (panel B of table 1 reports the summary statistics; fig. A.9*A*–A.9*N* plots 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

[18] The average difference between experimentation sites and nonexperimentation sites is 961 million yuan in terms of local GDP (26.0% of the average nonexperimentation 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.

preexperimentation characteristics and replicate the test for representativeness (average $t_{agriculture} = 2.48$, $t_{fiscal} = 5.26$, $t_{population} = 2.61$; see fig. A.10*A*–A.10*C*). We continue to find strong patterns of positive selection. For example, agricultural policy experiments take place in localities with substantially higher preexperimentation 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 populations. Third, pooling all policies together and focusing on preexperimentation fiscal expenditure in the policy-specific expenditure categories, we again find strong patterns of positive selection (see fig. A.10*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 fig. A.10*E*). Pro-rich policies indeed exhibit more positive site selection than pro-poor policies; however, even 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 fig. A.10*F*), by the total number of firms in the locality in the case of experiments with government finance and tax policies (see fig. A.10*G*), and by the total population size in the case of experiments with population and health policies (see fig. A.10*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 fig. A.11).

C. 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 III, we measure ex ante uncertainty for each policy experiment using proxies such as whether the central government has already laid out detailed national policy rollout plans at the beginning of the experiment and the level of consensus exhibited by academic publications before the experiment. Panel A of table A.3 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 rollout 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 III, 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. Panel B of table A.3 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 follows:

> 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 panel C of table 1). 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 VI.

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

VI. 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 IV, 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.

Columns 1–4 of table A.5 present the results. We find that, while participation in experiments is not per se 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 and 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.

A. 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

[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. Figure A.8 plots the overall share of positively selected experiments over the four decades since 1980, and table A.4 presents regression results on the time trend in positive selection, for all experiments and separately by ministry.

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} represents the ratio of fiscal domain k specific to the experiment in the total fiscal expenditure in county i during year t and Exp_{ikt} represents 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 allows 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 columns 1–3 of panel A in table 2. 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 (cols. 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]

This heterogeneity is consistent with the hypothesis that politically incentivized local leaders are particularly keen on making sure the policy

[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 zero or one.

[22] Local fiscal expenditure data (along with fiscal revenue) are among the least manipulable information due to their 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 term/position, his position, and his official rank in the bureaucratic system. See app. sec. B.1 for details.

TABLE 2
Local Fiscal Expenditure during Policy Experimentation

	Share of Fiscal Expenditure on Experimentation-Related Domains					
	(1)	(2)	(3)	(4)	(5)	(6)
	A. Fiscal Input among Experimentation Sites					
Number of experiments	.003*** (.001)	.002*** (.0004)	.002*** (.0005)	−.013*** (.003)	−.002* (.001)	−.002 (.002)
# × career incentive				.036*** (.006)	.009*** (.003)	.011*** (.003)
	B. Fiscal Input among Nonexperimentation Sites during National Policy Rollout					
Number of rolled-out policies	.001 (.001)	.001 (.0004)	.001 (.001)	.001 (.003)	.001 (.001)	.001 (.002)
# × career incentive				−.001 (.005)	−.0004 (.002)	−.0003 (.003)
Number of observations	150,977	150,977	150,977	142,128	142,116	142,116
Number of clusters	1,973	1,973	1,973	1,973	1,973	1,973
Mean of dependent variable	.173	.173	.173	.173	.173	.173
County by category fixed effects	No	Yes	Yes	No	Yes	Yes
Year by county fixed effects	Yes	No	Yes	Yes	No	Yes
Category by year fixed effects	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 reallocated 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; SD = 0.075). In panel B, we investigate whether the previously nonexperimentation sites exhibited similar fiscal reallocation in the year that the policy rolled out nationally. Standard errors (shown in parentheses) are clustered at the county level.

* $p < .10$.
*** $p < .01$.

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 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]

[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 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

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. Figure A.12 plots the yearly estimates 5 years before and 4 years after the start of the experiment (4 years is the average duration of experiments). We observe little evidence of a pretrend 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 table A.10). Second, we conduct the regression analysis at the policy-county level instead, controlling for experiment fixed effects, and thus exploiting variations in political incentives within the experiment across participating localities. We observe a consistent pattern: local politicians who have stronger career incentives are spending more fiscal resources during the experiment compared with other politicians participating in the same experiment (see 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 nonexperimentation 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 nonexperimentation sites (see panel B of table 2).[27] Moreover, among experimentation sites, increase in fiscal expenditure on the experimentation domain stops after the completion of experimentation (see 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

increases by 50%; see table A.8), reflecting a multitasking problem faced by the local politicians.

[26] Interestingly, local politicians in richer jurisdictions do not show stronger fiscal responses to policy experiments (see table A.9).

[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).

earn political credit as a "model experimentation site." 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 semantic analysis. 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 section G presents the details of the empirical specification and discussion of the results.

VII. 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 rollout, with section VII.A focusing on locality-specific shocks and section VII.B 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, regardless of its objective function. In the language of the framework presented in section IV, we test whether $\delta_2 \neq 1$.

A. 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* × *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° 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 predetermined 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 sites—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 (2SLS) specification:

$$LandRevenue_{ipt} = \alpha \cdot Suitability_i \times Interest_t + X'_{it}\beta + \delta_i + \gamma_t + \delta_m + \epsilon_{ipt},$$

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

where $LandRevenue_{ipt}$ represents 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 was eventually rolled out to the entire country, ψ_i represents a full set of county fixed effects, δ_m represents a full set of ministry fixed effects, and ν_t represents 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-statistic = 622.9; see table A.14). Panel A of table 3 presents the second-stage results. We find

[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.

[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 instrumental variables (IV) estimates.

TABLE 3
Naive Evaluation of Policy Experimentation

	National Rollout		
	(1)	(2)	(3)
	A. Land Revenue Windfall		
Land revenue (instrumented)	.008***	.006***	.009***
	(.002)	(.001)	(.001)
First-stage *F*-statistics	670.34	636.36	622.90
Number of observations	66,128	66,128	66,128
Number of clusters	1,644	1,644	1,644
Mean of DV	.612	.612	.612
Experiment year fixed effects	Yes	Yes	Yes
County fixed effects	No	Yes	Yes
Ministry fixed effects	No	No	Yes
	B. Politicians' Incentive Changes due to Political Rotation		
Rotation	.004	.016	.017
	(.019)	(.014)	(.014)
Positive rotation × ΔIncentive	.679***	.539***	.508***
	(.089)	(.090)	(.093)
Negative rotation × ΔIncentive	−.496***	−.459***	−.407***
	(.162)	(.132)	(.141)
Number of observations	3,899	3,899	3,899
Number of clusters	27	27	27
Mean of DV	.321	.321	.321
Ministry fixed effects	No	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Province fixed effects	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 (shown in parentheses) are clustered at the county level. We report the first-stage results in 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 2 years after the last round. An average positive rotation is accompanied with an incentive increase of 0.079 (SD = 0.076). An average negative rotation is accompanied by an incentive drop of 0.055 (SD = 0.061). The standard errors are clustered at the province level.

*** $p < .01$.

robust positive coefficients of instrumented land revenue at experimentation sites on the corresponding policy's national rollout.[30] In other words, when policy experimentation is conducted in localities that coincidentally

[30] Our baseline analysis is conducted at the county-policy-year level—if a county has two ongoing policy experiments in a given year, they show 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

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 rollout decisions are more affected by land revenue windfalls in experiments with fewer participating sites (see table A.16), consistent with the fact that each locality plays a larger 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 autocorrelation in interest rates). As shown in 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—the ratio of land between 15° and 30° × contemporaneous national interest rate—affects local land revenue. Since the government's official cutoff for real estate development is 15°, only land plots with slopes below this cutoff would matter for real estate construction. As shown in panel B of table A.17, this is indeed the case: land plots between 15° and 30° 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 × national interest rate at $t + 5$ as the instrument for land revenue at $t + 5$ (since the vast majority of policy experiments conclude within 5 years). As we can see in panel C of table A.17, while there is a very strong first stage, land revenue at $t + 5$ has a precisely estimated null effect on the rollout of policies being experimented in year t.

B. *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?

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 table A.15, our findings are robust to this alternative way of structuring the data.

In particular, we examine whether changes in local politicians' career incentives (and thus changes in effort, as shown in sec. VI) due to local politicians' routine turnover affect the central government's policy learning and increase 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:

$$\begin{aligned} y_p = {} & \alpha \cdot \mathit{Turnover}_{ip} + \beta_1 \cdot \mathit{Turnover}_{ip} \times \mathit{IncreaseIncentive}_{ip} \\ & + \beta_2 \cdot \mathit{Turnover}_{ip} \times \mathit{DecreaseIncentive}_{ip} + \gamma_t + \delta_m + \theta_n + \varepsilon_{ipmnt}, \end{aligned}$$

where y_p is the indicator of experiment p being evaluated as successful and rolled out to the entire country and $\mathit{Turnover}_{ip}$ 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 $\mathit{Incentive}_{ip}$ 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 app. sec. B.1). We separate local political turnovers that result in either an increase ($\mathit{IncreaseIncentive}_{ip}$) or a decrease ($\mathit{DecreaseIncentive}_{ip}$) 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.

Panel B of table 3 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 rollout. However, when the incoming politician has stronger upward career mobility potential than the outgoing politician (i.e., younger vs. 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 rollout decisions is more pronounced among small-scale experiments (see 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 cutoff (see panel A of table A.19). 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 5 years after the beginning of the experimentation period (see panels D.1 and D.2, respectively).

VIII. What Are the Implications on Policy Learning and Policy Outcomes?

Having documented three facts about China's policy experimentation in sections V–VII, we now examine their implications for the central government's policy learning and the effectiveness of national policies originating from such experimentation. In section VIII.A, we discuss such implications under the assumption that the central government of China aims to learn about policies' ATE. In section VIII.B, we discuss alternative objectives of policy experimentation beyond learning about policies' ATE.

A. If Experimentation Objective Is to Learn about Policies' ATE

1. Experimentation Outcomes and Policies' National Rollout

While we do not directly observe how the central government evaluates policy experiments and decides on policies' rollout, we can infer the decision rule by examining which estimators of experimentation outcomes most strongly predict the corresponding policies' national rollout. We begin with a simple estimator of experimentation effects that compares

[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.

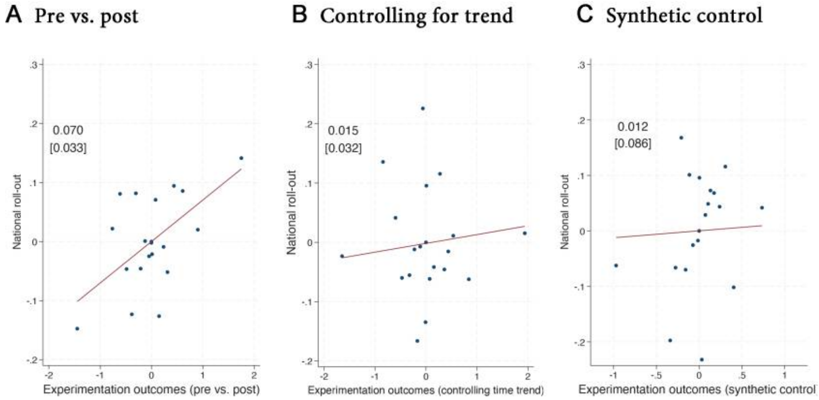


FIG. 4.—These plots visualize the correlations between policy rollout rate and different measures of experimentation effects. In *A*, we construct the simple difference in GDP per capita, among the experiment sites, before and after the policy trial; in *B*, we further control for provincial pretrends in GDP per capita; in *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 pretrend, and minister and year fixed effects are included. Coefficients and robust standard errors are reported in the top left corner of each panel.

experimentation sites' local economic performance before and after the experiments, averaged across all experimentation sites. Figure 4*A* presents the correlation between the estimated experimentation effects (on the *x*-axis) and the decision to roll out a policy nationally (on the *y*-axis). There is a strongly positive correlation between the two: a 1 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 panel A of table A.20).[32]

[32] In table A.21, we reestimate the correlation between the estimated experimentation effects and the national rollout decision, winsorizing the sample by 2.5% 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 sec. VIII.B.1. The sample for this analysis starts in 1993, which was when county-level socioeconomic data started to be reliably published in the Statistical Yearbooks.

The comparison of experimentation sites' economic performance before and after the experiments is not informative of the experimental policy's ATE, since it does not account for positive site selection and nonrepresentative 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 nonexperimentation sites based on 5-year preexperimentation trends in local socioeconomic conditions. The correlations between these estimated experimentation effects and policies' national rollout are plotted in fig. 4*B* and 4*C*, respectively (and in panels B and C of table A.20, in regression form). These more sophisticated estimates of experimentation effects, which would have been more informative of policies' ATE, no longer predict whether policies roll out nationwide.

Assessing the magnitude.—Leveraging the estimates from sections V–VII, we perform a back-of-the-envelope calculation on how much the national policy rollout 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 (for details, see app. sec. H).[33] A 1% increase in fiscal revenue for all experimentation sites would increase the corresponding policy's national rollout probability by 1.8 percentage points. Linking this number to the average (preexperimentation) difference in fiscal revenue between experimentation and nonexperimentation sites (20.5%), we calculate that positive site selection inflates the national rollout 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 nonrepresentative fiscal resources inflate the national rollout rate of an average policy experiment by an additional 24.1%.[34]

[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, nonexperimentation 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 nonexperimentation 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 sec. VII.A, a 1% increase in local politicians' promotion incentives would increase the corresponding policy's national rollout probability by 0.68 percentage points. Linking this elasticity to the average difference in local politicians' incentives between experimentation and nonexperimentation sites (1.3%), we calculate that nonrepresentative political incentives inflate the national rollout rate of an average policy experiment by an additional 2.2%.

2. National Policy Outcomes

Both positive experimentation site selection and extra efforts among local politicians during the experiment 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 subsection, when constructing measures of policy outcomes during either experimentation or rollout, 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 with experimentation outcomes?.— We begin by examining the potential shrinkage of experimentation effects across all experimental economic policies. In figure 5*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 nonexperimentation sites before and after the corresponding policies roll out to the country). Figure 5*B* plots the distribution of the differences of experimentation and national policy effects. One observes that many policies (71.1%) fall below the 45° line, reflecting smaller effects during the national rollout.[36] In fact, while the (naively estimated) experimentation effects strongly predict policies' national rollout, 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 focus on policy effects on local fiscal revenue, an indicator of local economic performance that is unlikely to be manipulated by the local politicians (see figs. A.16*A*, A.17*A*), and (ii) correct for local economic growth (mis)reporting using local

[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 fig. 5, controlling for the number of experimentation sites (namely, the sample size for each experiment). The results are presented in fig. 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 fig. 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 nonexperimentation sites increase, the experimentation effects become substantially less predictive of the national policy effects.

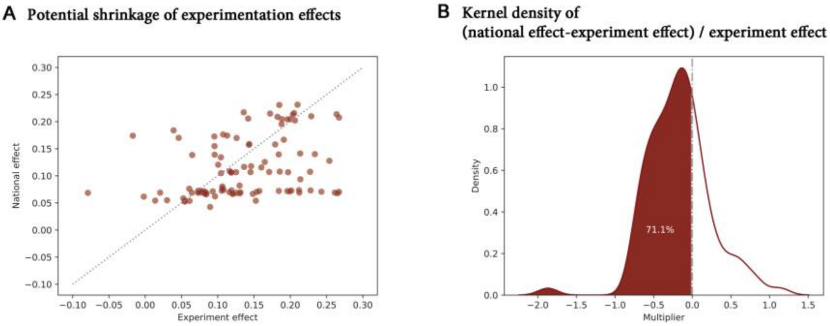


Fig. 5.—These plots demonstrate how policy effects shrink between the experimentation and rollout stages. In *A*, we plot policy effect during national rollout (*y*-axis) against experimentation effect of the same policy (*x*-axis). In *B*, we compute the difference between policy effect during national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

luminosity from satellite images, following Martinez (2022; see figs. A.16*B*, A.17*B*).

The shrinkage of experimentation effects as a policy rolls out could result from a combination of the lack of representativeness of experimentation sites (in terms of both 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 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. To foster economic growth, the central government initiated an experiment that provides more fiscal autonomy to the counties participating in the experiment (for policy details, see app. sec. A.2). Between 2003 and 2013, more than 1,100 counties were selected as experimentation sites. The

[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 fig. A.18).

experimentation sites were positively selected during the first half of the experiment and moved to negative selection toward the end of the experiment (t-statistic > 10 in 2004, t-statistic < -3 in 2017; for more details, see fig. A.20). We use a staggered event study design to estimate the treatment effects of the introduction of such a 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 preexperimentation GDP per capita benefited from the experiment, while the poorer counties experienced worse subsequent local economic development (see fig. A.21).[39] The local fiscal empowerment experiment 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 fig. 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 preexperimentation 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 originated from:

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

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

[39] Such patterns of heterogeneity by preexperimentation local economic conditions do not merely reflect a general equilibrium effect or an early-mover advantage in reform. Less developed counties participating in the experiment during the early rounds also experienced a negative policy treatment effect in magnitudes similar to the less developed experimentation sites in later rounds (see fig. A.22).

TABLE 4
Similarity with Experimentation Sites and Effects of Policy Rollout

	Growth of GDP per Capita		
	(1)	(2)	(3)
	A. Selection of Experimentation Sites		
Mahalanobis distance in socioeconomic conditions	−.004*** (.001)	−.004*** (.001)	−.003*** (.001)
Number of observations	94,772	94,772	94,772
Number of clusters	2,064	2,064	2,064
Mean of DV	.102	.102	.102
	B. Endogenous Efforts during Experimentation		
Mahalanobis distance in politicians' career incentives	−.001*** (.0002)	−.001*** (.0003)	−.0002 (.0002)
Number of observations	55,940	55,940	55,940
Number of clusters	1,464	1,464	1,464
Mean of DV	.088	.088	.088
Policy fixed effects	No	No	Yes
Year fixed effects	No	Yes	Yes
County fixed effects	Yes	Yes	Yes

Note.—This table investigates how much of a policy's effectiveness at the national rollout stage can be attributed to the site selection and endogenous effort patterns at its experimentation stage. The sample includes all nonexperimentation 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 nonexperimentation counties for a given policy experiment, in terms of their socioeconomic conditions. In panel B, we investigate Mahalanobis distance between the experimentation and nonexperimentation 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 app. sec. 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 (shown in parentheses) are clustered at the county level.

*** $p < .01$.

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, nonexperimentation 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 indexes chosen to compute the distance (see 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 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 interregional inequality throughout China.

B. Alternative Experimentation Objectives

In section VIII.A, 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 ATE 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.

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 rollout of the corresponding experimental policy. Figure A.24 presents the estimated coefficients across the percentile thresholds, which range from zero to the 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 ATE (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 unrepresentative 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 section 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 section 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 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.

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-statistics 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 intertemporal 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 figure A.25, we plot the event study estimates around ministers' turnover; reassuringly, we do not observe a pretrend.

As shown in panel A of table A.23, 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 driven almost entirely 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 the Global Database of Events, Language, and Tone, 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 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 rollout. 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, measured as in section VIII.A.1, unrest episodes are associated with substantially lower chances that the local experimental policies would eventually become national policies (see panel A of table A.25). 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 rollout, we follow Beraja et al. (2023) and use local weather conditions to instrument for protest occurrence.[42] As shown in panel B of table A.25, weather-induced variations in protests strongly predict the national rollout 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 rollout.

IX. 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 rollout, if the central government intends to learn about the policy's average effects in a representative locality with a 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 bureaucratic environment could affect the initiation of policy experimentation, its structure and implementation, and its 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

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

Rodrik 2005; Callander and Harstad 2015; Myerson 2015).[43] Rather than the informational free-riding and underexperimentation observed in federalist systems, political centralization—in a context such as China where local government officials compete and differentiate their implementation activities to increase their chances of promotion—could overcome tendencies of underexperimentation.[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 to stimulate politicians' efforts to improve policy outcomes, while making sure that such incentives are not exaggerated during the experimentation phase, so that policy learning remains unbiased. Dynamic experimentation could be a solution (e.g., Kasy and Sautmann

[43] However, Cheng and Li (2019) note that the uncertainty related to citizens' inference on politicians' types could induce politicians to overexperiment even in a decentralized environment.

[44] Following Goldszmidt et al. (2020), Holz et al. (2020), and List (2020), we examine the SANS (selection, attrition, naturalness, and scaling) 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 policymaking in China for 40 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 "nonnegotiable" 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 to achieve better policymaking.

[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).

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. For example, this study does not examine 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 policymaking cycle would be a fascinating, important, and challenging undertaking that we leave for future work.

Data Availability

Data and code replicating the tables and figures in this article can be found in the Harvard Dataverse, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/38SN2Y (Wang and Yang 2024).

References

Aghion, Philippe, Patrick Bolton, Christopher Harris, and Bruno Jullien. 1991. "Optimal Learning by Experimentation." *Rev. Econ. Studies* 58 (4): 621–54.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Q.J.E.* 130 (3): 1117–65.

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.

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." *J. Econ. 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." Working Paper no. 25848, NBER, Cambridge, MA.

Bai, Chong-En, Chang-Tai Hsieh, and Zheng Song. 2020. "Special Deals with Chinese Characteristics." *NBER Macroeconomics Ann.* 34 (1): 341–79.

Banerjee, Abhijit V., Sylvain Chassang, Sergio Montero, and Erik Snowberg. 2020. "A Theory of Experimenters: Robustness, Randomization, and Balance." *A.E.R.* 110 (4): 1206–30.

Beraja, Martin, Andrew Kao, David Y. Yang, and Noam Yuchtman. 2023. "AI-tocracy." *Q.J.E.* 138 (3): 1349–402.

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, Univ. California, Berkeley.

Bertrand, Marianne, Matilde Bombardini, Raymond Fisman, Brad Hackinen, and Francesco Trebbi. 2020. "Hall of Mirrors: Corporate Philanthropy and Strategic Advocacy." Tech. report, Dept. Econ., Boston Univ.

Blanchard, Olivier, and Andrei Shleifer. 2001. "Federalism With and Without Political Centralization: China versus Russia." *IMF Staff Papers* 48 (1): 171–79.

Cai, Hongbin, and Daniel Treisman. 2009. "Political Decentralization and Policy Experimentation." *Q. J. Polit. Sci.* 4 (1): 35–58.

Callander, Steven. 2011. "Searching for Good Policies." *American Polit. Sci. Rev.* 105 (4): 643–62.

Callander, Steven, and Bård Harstad. 2015. "Experimentation in Federal Systems." *Q.J.E.* 130 (2): 951–1002.

Cao, Yuanzheng, Yingyi Qian, and Barry R. Weingast. 1999. "From Federalism, Chinese Style to Privatization, Chinese Style." *Econ. Transition* 7 (1): 103–31.

Chandler, Alfred Dupont. 1962. *Strategy and Structure: Chapters in the History of the Industrial Enterprise*, vol. 120. Cambridge, MA: MIT Press.

Chen, Ting, and J. K.-S. Kung. 2016. "Do Land Revenue Windfalls Create a Political Resource Curse? Evidence from China." *J. Development Econ.* 123:86–106.

Cheng, Chen, and Christopher Li. 2019. "Laboratories of Democracy: Policy Experimentation under Decentralization." *American Econ. J.: Microeconomics* 11 (3): 125–54.

Davis, Jonathan M. V., Jonathan Guryan, Kelly Hallberg, and Jens Ludwig. 2017. "The Economics of Scale-Up." Working Paper no. 23925, NBER, Cambridge, MA.

DellaVigna, Stefano, and Woojin Kim. 2022. "Policy Diffusion and Polarization across U.S. States." Working Paper no. 30142, NBER, Cambridge, MA.

DellaVigna, Stefano, and Elizabeth Linos. 2020. "RCTs to Scale: Comprehensive Evidence from Two Nudge Units." Working Paper no. 27594, NBER, Cambridge, MA.

Dewatripont, Mathias, and Gerard Roland. 1995. "The Design of Reform Packages under Uncertainty." *A.E.R.* 85 (5): 1207–23.

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." Working Paper no. 27628, NBER, Cambridge, MA.

Fisman, Raymond, Jing Shi, Yongxiang Wang, and Weixing Wu. 2020. "Social Ties and the Selection of China's Political Elite." *A.E.R.* 110 (6): 1752–81.

Fisman, Raymond, and Yongxiang Wang. 2015. "The Mortality Cost of Political Connections." *Rev. Econ. Studies* 82 (4): 1346–82.

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

Glaeser, Edward L., and Andrei Shleifer. 2003. "The Rise of the Regulatory State." *J. Econ. Literature* 41 (2): 401–25.

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." Working Paper no. 28208, NBER, Cambridge, MA.

Hayek, Friedrich August. 1978. *Rules and Order*, vol. 1 of *Law, Legislation, and Liberty*. Chicago: Univ. Chicago Press.

He, Guojun, Shaoda Wang, and Bing Zhang. 2020. "Watering Down Environmental Regulation in China." *Q.J.E.* 135 (4): 2135–85.

Heilmann, Sebastian. 2008a. "From Local Experiments to National Policy: The Origins of China's Distinctive Policy Process." *China J.* 59:1–30.

———. 2008b. "Policy Experimentation in China's Economic Rise." *Studies Comparative Internat. Development* 43 (1): 1–26.

Hirsch, Alexander V. 2016. "Experimentation and Persuasion in Political Organizations." *American Polit. Sci. Rev.* 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." *A.E.R.* 111 (5): 1442–80.

Holz, Justin E., John A. List, Alejandro Zentner, Marvin Cardoza, and Joaquin Zentner. 2020. “The $100 Million Nudge: Increasing Tax Compliance of Businesses and the Self-Employed Using a Natural Field Experiment.” Working Paper no. 27666, NBER, Cambridge, MA.

Jia, Junxue, Qingwang Guo, and Jing Zhang. 2014. “Fiscal Decentralization and Local Expenditure Policy in China.” *China Econ. Rev.* 28:107–22.

Jia, Ruixue, Masayuki Kudamatsu, and David Seim. 2015. “Political Selection in China: The Complementary Roles of Connections and Performance.” *J. European Econ. Assoc.* 13 (4): 631–68.

Jiang, Junyan. 2018. “Making Bureaucracy Work: Patronage Networks, Performance Incentives, and Economic Development in China.” *American J. Polit. Sci.* 62 (4): 982–99.

Johnson, Simon, and James Kwak. 2011. *13 Bankers: The Wall Street Takeover and the Next Financial Meltdown.* New York: Vintage.

Kasy, Maximilian, and Anja Sautmann. 2021. “Adaptive Treatment Assignment in Experiments for Policy Choice.” *Econometrica* 89 (1): 113–32.

Kornai, Janos. 1959. *Overcentralization in Economic Administration: A Critical Analysis based on Experience in Hungarian Light Industry*. London: Oxford Univ. Press.

Lan, Xiaohuan. 2021. *Embedded Power: Chinese Government and Economic Development.* Shanghai: Shanghai People’s Publishing House.

Li, Hongbin, and Li-An Zhou. 2005. “Political Turnover and Economic Performance: The Incentive Role of Personnel Control in China.” *J. Public Econ.* 89 (9/10): 1743–62.

List, John A. 2020. “Non est Disputandum de Generalizability? A Glimpse into the External Validity Trial.” Working Paper no. 27535, NBER, Cambridge, MA.

———. 2022. *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale.* New York: Currency.

Martinez, Luis R. 2022. “How Much Should We Trust the Dictator’s GDP Growth Estimates?” *J.P.E.* 130 (10): 2731–69.

Mehmood, Sultan, Shaheen Naseer, and Daniel L. Chen. 2021. “Training Policymakers in Econometrics.” Working paper.

Montinola, Gabriella, Yingyi Qian, and Barry R. Weingast. 1995. “Federalism, Chinese Style: The Political Basis for Economic Success in China.” *World Polit.* 48 (1): 50–81.

Mukand, Sharun W., and Dani Rodrik. 2005. “In Search of the Holy Grail: Policy Convergence, Experimentation, and Economic Performance.” *A.E.R.* 95 (1): 374–83.

Muralidharan, Karthik, and Paul Niehaus. 2017. “Experimentation at Scale.” *J. Econ. Perspectives* 31 (4): 103–24.

Myerson, Roger. 2015. “Local Agency Costs of Political Centralization.” Working paper, Univ. Chicago.

Narita, Yusuke. 2021. “Incorporating Ethics and Welfare into Randomized Experiments.” *Proc. Nat. Acad. Sci.* 118 (1): e2008740118.

Naughton, Barry. 1996. *Growing Out of the Plan: Chinese Economic Reform, 1978–1993.* Cambridge: Cambridge Univ. Press.

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance.* Cambridge: Cambridge Univ. Press.

Qian, Yingyi. 2002. “How Reform Worked in China.” Working paper.

Qian, Yingyi, Gerard Roland, and Chenggang Xu. 2006. “Coordination and Experimentation in M-Form and U-Form Organizations.” *J.P.E.* 114 (2): 366–402.

Rawski, Thomas G. 1995. “Implications of China’s Reform Experience.” *China Q.* 144:1150–73.

Rogger, Daniel, and Ravi Somani. 2018. “Hierarchy and Information.” Working Paper no. 8644, World Bank, Washington, DC.

Roland, Gerard. 2000. *Transition and Economics: Politics, Markets, and Firms.* Cambridge, MA: MIT Press.

Sachs, Jeffrey D. 2006. *The End of Poverty: Economic Possibilities for Our Time.* New York: Penguin.

Shipan, Charles R., and Craig Volden. 2006. “Bottom-Up Federalism: The Diffusion of Antismoking Policies from US Cities to States.” *American J. Polit. Sci.* 50 (4): 825–43.

Shleifer, Andrei. 1996. “Origins of Bad Policies: Control, Corruption and Confusion.” *Rivista Polit. Econ.* 108 (3): 599–617.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020. “Specification Curve Analysis.” *Nature Human Behaviour* 4 (11): 1208–14.

Snowberg, Erik, and Leeat Yariv. 2018. “Testing the Waters: Behavior across Subject Pools.” Working Paper no. 24781, NBER, Cambridge, MA.

Stigler, George J. 1971. “The Theory of Economic Regulation.” *Bell J. Econ. and Management Sci.* 2 (1): 3–21.

Vivalt, Eva. 2020. “How Much Can We Generalize from Impact Evaluations?” *J. European Econ. Assoc.* 18 (6): 3045–89.

Vivalt, Eva, and Aidan Coville. 2019. “How Do Policymakers Update?” Working paper.

Wang, Shaoda, and David Yang. 2024. “Replication Data for: ‘Policy Experimentation in China: The Political Economy of Policy Learning.’” Harvard Dataverse, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/38SN2Y.

Wang, Zhi, Qinghua Zhang, and Li-An Zhou. 2020. “Career Incentives of City Leaders and Urban Spatial Expansion in China.” *Rev. Econ. and Statis.* 102 (5): 897–911.

Williamson, Oliver E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications: A Study in the Economics of Internal Organization.* New York: Free Press.

Xie, Yinxi, and Yang Xie. 2017. “Machiavellian Experimentation.” *J. Comparative Econ.* 45 (4): 685–711.

Xu, Chenggang. 2011. “The Fundamental Institutions of China’s Reforms and Development.” *J. Econ. Literature* 49 (4): 1076–151.

Xu, Yiqing. 2017. “Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models.” *Polit. Analysis* 25 (1): 57–76.

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