The Political Economy of Policy Experimentations in China

Shaoda Wang
David Y. Yang*

September 1, 2021

Abstract

Many governments have engaged in policy experimentations in various forms to resolve policy uncertainty and to facilitate policy learning. However, little is understood about the characteristics of policy experimentations, and how the structure of experimentations may affect policy learning and policy outcomes. In this project, we aim to describe and understand China's policy experimentations since the 1980s, among the largest and most systematic in recent history. We collect comprehensive data on policy experimentations conducted in China over the past 4 decades by 98 ministries and commissions. We find three main results. First, more than 80% of the experimentations exhibit positive sample selection in terms of locality's economic development, and much of the positive selection can be attributed to misaligned incentives between central and local government. Second, local politicians exert greater effort and allocate more resources to ensure the success of policy experimentations, often at the expense of crowding out efforts allocated to other domains. Third, the presence of sample selection and strategic effort is not fully accounted for by the central government, thus affecting policy learning and distorting national policies originated from experimentations. Taken together, these results suggest that policy experimentations take place under various political and bureaucratic constraints, which could limit the scope and bias the direction of policy learning.

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

Crossing the river by touching the stones.

Chen Yun

1 Introduction

Policy effectiveness often hinges on the nature of policy in consideration, its implementation details, degree of tailoring to local conditions, as well as efforts and incentives of local politicians to make the policy work. Determining which policies to implement and how to implement them is an essential component of government's decision making (e.g., Hayek, 1978; North et al., 1990).

Many governments have explicitly or implicitly engaged in policy experimentations in various forms in order to resolve policy uncertainty and to facilitate policy learning (e.g., Roland, 2000; Mukand and Rodrik, 2005). Sophisticated policy experimentations range from sequences of trials and errors, to rigorous randomized control trials in subregions of a country. Few, however, can compare to the systematic policy experimentations in China in terms of their breadth, depth, and duration. Since the 1980s, the Chinese government has been trying out different policies across regions and often over multiple waves before deciding on rolling out the policies to the entire nation. Scholars have argued that the pursuit of extensive, continuous, and institutionalized experimentations are critical mechanism that leads to China's economic rise and gradual reform over the past four decades (e.g., Rawski, 1995; Cao et al., 1999; Roland, 2000; Qian, 2002).

However, little is understood about the characteristics of policy experimentations, and how the structure of experimentations may affect policy learning and policy outcomes. Policy experimentations could provide important and informative signals on policy effectiveness, particularly if the following two conditions are satisfied. First, to the extent that policy effects are often heterogeneous across localities, representative selection of experimentations sites is critical to ensure unbiased learning of the policy's average effects. Second, to the extent that efforts of the key actors (such as local politicians) may play important role in shaping policy outcomes, experimentations that do not induce additional incentives and excessive efforts could ensure that signals of policy effectiveness during the experimentations are not exaggerated.

This project aims to describe and understand China's policy experimentations since the 1980s. Specifically, we aim to answer three questions. First, is the sample selection of China's policy experimentations representative? Second, do policy experimentation

induce additional incentives and efforts ("experimenter effects") that could exaggerate policy effectiveness? And third, how do the sample selection and experimenter effects affect government's policy learning as well as national policy outcomes?

To answer these questions, we collect comprehensive data on policy experimentations in China between 1980 and 2020. Based on 19,812 government documents, we construct a database of 663 policy experimentations initiated by 98 central ministries and commissions. For each policy experimentation, we link the central government document that outlines the overall experimentation guidelines with all corresponding local government documents to record its local implementation, and trace its roll-out across the country. We measure a variety of characteristics of policy experimentations based on the associated government documents and other linked datasets, including the ex-ante uncertainty of the policy at trial, career trajectories of central and local politicians involved in the experimentations, the bureaucratic structure of the policy-initiating ministries, and local socioeconomic conditions.

We begin with investigating the selection of experimentation sites. Being able to learn from policy experimentations from a balanced, representative sample is a primary goal for the central government, as prescribed by the the National Development and Reform Commission who oversees many key experimentations. Nonetheless, comparing the pre-experimentation characteristics between the locations that are selected as experimentation sites and those that are not, we observe that more than 80% of the experimentations are conducted in sites that are positively selected in terms of local economic conditions.

What explains such deviation from representativeness? A number of findings suggest that such positive selection cannot be justified merely out of optimal experimentation considerations.[1] For example, experimentation sites for policies that are ex-ante more uncertain are *more* positively selected, suggesting that positive selection is not merely a result of objectives unrelated to resolve policy uncertainty. Experimentations that are more complicated (e.g., involving coordination across multiple ministries) are *not* substantially more positively selected than simpler experimentations, suggesting that positive selection is not driven by the possibility that richer places may provide more precise signals for complicated policies. Moreover, experimentations in domains typically associated with national and uniform policies are *more* positively selected than those that are more flexible for regional targeting, suggesting that deviation from representativeness is unlikely due to the regional scope of certain policies.

[1]While deviation from randomization could be justified under alternative incentives such as maximizing decision-makers' subjective expected utility and experimentation sites' welfare, our simulations of the optimal experimentation design demonstrate that these incentives alone can only account for a very small fraction of the degree of positive selection that we observe.

In fact, we find that nearly half of the positive selection can be accounted for by misaligned incentives between central government who desires experimentations to generate nationally-representative knowledge about policy effectiveness, and the local governments who aim to demonstrate trial success out of career concerns.[2] We document two specific political economy factors that may limit the central government's ability to achieve representativeness. First, local government officials who face stronger career incentives (due to a combination of mandatory retirement age and room for upward mobility in the career trajectory) are significantly more likely to engage in policy experimentations. Second, the selection of experimentation sites is affected by political patronage: local politicians who are connected to the ministers who oversee the experimentations are substantially more likely to have their jurisdictions selected as experimentation sites. Both of these patterns suggest that even if the central government's only objective is to obtain a representative sample of experimentation sites, deviation from representativeness could occur when ministers and local governments implement the policy experimentations.

Next, we examine whether policy experimentations induce "experimenter effects" in two important dimensions of politicians' strategic efforts during experimentations — fiscal resource allocation and attempts to differentiate in implementation details. First, using a triple-differences strategy, we find that during experimentations, local governments incur almost 5% more fiscal expenditure in the categories relevant to the policy at trial, and this is particularly the case for experimentations that do not involve explicit fiscal support from the central government. Such increase in fiscal support comes in part at a cost of crowding out expenditure in other fiscal categories unrelated to the ongoing experimentations. Second, we find that among local politicians participating in a specific policy experimentation, those facing greater career incentives (again due to age before retirement and room for upward mobility) are significantly more likely to differentiate in policy implementation details, captured by pairwise text similarity among the governments' experimentation documents. Such differentiation could earn political credits for local politicians, such as the selection of "model experimentation sites" by the central government.

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

[2]The central government holds other criteria in choosing experimentation sites that may at times counteract its desire for representativeness. Specifically, we find that prefectures that have experienced social and political unrest in the preceding period are substantially less likely to be assigned as experimentation sites, suggesting that unstable local environment could be a veto condition that precludes experimentations. Such multi-dimensional objectives of the central government could further push the experimentation site selection away from being representative.

ernment's policy learning and the national policy outcomes. We present evidence that the central government does not fully account for sample selection and strategic effort when evaluating policy experimentations: in particular, successful local trials due to positive, exogenous shocks to fiscal resources and political incentives much more likely to become national policies. We explore two sources of exogenous shocks. First, experimentations are substantially more likely to be rolled out to the entire country if they take place among localities experiencing greater land revenue windfall due to abundance of land suitable for commercial and real estate development as well as time-varying local demand shocks during the experimentations. Such land revenue increase constitutes a considerable portion of local fiscal capacity, and thus could affect economic performance among the experimentation sites and provide resources that enhance many local policy outcomes.[3] Second, while local politicians with high career incentives differentially exert efforts during policy experimentations, an increase in career incentives due to political transitions *after* the experimentations start should be irrelevant to the underlying policy effectiveness and ought to be discarded by central government's policy inference. Nonetheless, we find that after a particular experimentation has started, if local government transitions to politicians with stronger career incentives, such experimentation is significantly more likely to become successful and rolled out to the entire country.

Evaluations of experimentation outcomes in the presence of positive sample selection and experimenter effects can influence national policy outcomes. When experimented policies roll out to the entire country, localities that shared similar socioeconomic conditions with the sites where the corresponding policies were experimented on (often richer localities) benefit substantially more from the policies. This could systematically bias the space of implemented policies and reforms in China, and generate distributional consequences across regions.

Taken together, these results demonstrate that China's remarkable policy experimentations, as with any other undertaking in policy learning at this scale, cannot be fully understood without taking into account the complex institutional (both political and organizational) conditions and contexts. On the one hand, a set of the institutional and bureaucratic conditions may serve as the engine to coordinate experimentations, to motivate politicians' participation, and to stimulate local policy innovations. Experimentations thus can help circumvent political and bureaucratic frictions that may prevent reform and policy adoption. On the other hand, as we have demonstrated, the very same institu-

[3]More generally, we observe that experimentations that are closer to representativeness in their sites selection are *less* likely to become formal, national policies, consistent with the pattern that representative experimentations are more likely to reveal sub-optimal policies given the uncertainty and heterogeneity in the mapping from policies to outcomes.

tional and bureaucratic contexts also imply the presence of constraints, such as incentive misalignment within the polity, which could result in deviation from representative sample selection and generate experimenter effects. To the extent that these characteristics of the policy experimentations are not sufficiently accounted for, policy learning can be biased and national policy outcomes may be affected.

This paper contributes to the (largely theoretical) literature on policy learning and policy experimentation. For example, Aghion et al. (1991) and Callander (2011) provide theoretical frameworks on searching (and experimenting) with good policies; Dewatripont and Roland (1995) provide justification for the experimental approach in policy reforms; Qian et al. (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; Callander and Harstad (2015) investigate how decentralized jurisdictions strategically engage in policy experimentation, and how a central government could help encourage policy convergence. Closest to the context we study, Montinola et al. (1995), Cao et al. (1999), Heilmann (2008a,b) and Xie and Xie (2017) study the institutional setup and political logic of China's policy experimentations. We contribute to these works by providing the first empirical analyses of the comprehensive set of policy experimentations that are conducted in China over the past four decades, linking the theoretical predictions on when policy experimentations should take place, how they should be structured, with the specific institutional contexts that may constrain such experimentations.[4]

Our work joins the recent empirical literature on policy learning and potential biases. The patterns of positive experimentation sites selection in general, and of the diminishing policy effects as the fiscal empowerment experimentation expands its roll-out echo the similar findings documented by Allcott (2015) on the sample selection bias in the Opower energy conservation programs in the US. Moreover, as we document that the Chinese government at times fails to disentangle factors not associated with inherent policy effectiveness when evaluating outcomes of policy experimentations, we also add to recent works highlighting that key policymakers may not be fully sophisticated at processing information (Rogger and Somani, 2018; Hjort et al., 2019; Vivalt and Coville, 2019).

Our analyses of policy experimentations in China also contribute to the political economy literature on federalism and political (de)centralization. Persson et al. (2000) and Besley and Coate (2003) provide seminal frameworks for analyzing the interaction be-

[4]In so doing, we also contribute to the large literature on the political economy behind China's spectacular growth. A large literature studies the Chinese economy and its spectacular growth in the recent decades (e.g., Song et al., 2011; Khandelwal et al., 2013; Roberts et al., 2018; Cheng et al., 2019), to which we bring empirical description of policy experimentations, often speculated as a key driver of China's growth.

tween the central and local governments, and Myerson (2015) represents recent progresses in understanding the agency costs associated with the degree of centralization in a political regime. We demonstrate, in the context of central government's designing and local government's implementation of policy experimentations, a misalignment between the central and local governments that is inherent with decentralization and the principal-agent problems. In so doing, we also bring together the political economy literature with the literature on organizational economics (e.g., Maskin et al., 2000), as we show that the designing of organizational structure of political institutions could have a considerable impact on the degree to which policy experimentations and policy innovations are carried out. Specifically, concerning the relationships between policy learning and political centralization, our findings of the patterns in policy experimentations in a centralized environment stand in contrast with theoretical works analyzing experimentations in decentralized, voluntary contexts (Mukand and Rodrik, 2005; Callander and Harstad, 2015).[5] While misaligned incentives between the central and local governments generates suboptimal learning, rather than informational free-riding and under-experimentation, such misalignment can induce over-experimentation in a centralized context such as China, where local governments compete and overly differentiate in order to increase promotion chances.[6]

Our findings that policies originated from unrepresentative experimentations could disproportionately benefit richer regions also contribute to the literature on regulatory capture and distortions in rule-making processes. In addition to the 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 philanthropy giving and strategic advocacy (Bertrand et al., 2020). Our results in the context of policy experimentations demonstrate yet another manifestation of regulatory capture by systematically biasing the information that the decision makers

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

[6]Relatedly, the existing literature has attributed China's success with economic decentralization to its high-powered political centralization (Blanchard and Shleifer, 2001; Xu, 2011), which fosters promotion competition among local politicians on dimensions aligned with the central government's policy goal (e.g., Li and Zhou, 2005; Jia et al., 2015; Bai et al., 2020). Our results complement this literature by illustrating that when the central government rewards local officials based on an incomplete contract (e.g., local officials get political rewards for delivering successful experimentations, without fully taking into account the representativeness and external validity issues), the local officials' behaviors could deviate from the central government's original intention in costly ways, highlighting a classic pitfall of political centralization (Kornai, 1959).

could gather during the policy learning process.[7]

Finally, our paper also connects to the large literature on experimental methods, in particular policy evaluations in the field. An emerging literature studies optimal experimentation design when sample sizes are finite and decision makers care about the welfare of the experiment participants in addition to pure learning objectives, which could justify deterministic experimentation (Banerjee et al., 2020; Narita, 2021). We confirm that in environments with finite sample and large regional heterogeneity, initial positive selection could be optimal, deterministic experimentation and some degree of initial positive selection of the experimentation sites may be justified. However, one needs to adjust the stopping rule of the experimentations in order to prevent policies that are experimented with non-representative sample to prematurely get rolled out nationally (see Kasy and Sautmann (2021) for a recent example of dynamic experimental design). Moreover, a recent literature examines the problems related to scaling up policy interventions in particular localities, emphasizing the general equilibrium consequences arising from such scaling up (e.g., Bergquist et al., 2019), and changes in the intervention when scaled up (DellaVigna and Linos, 2020). As we examine the consequences of deviation from representativeness in China's policy experimentations, we note that the underlying heterogeneous treatment effects and the related external validity are key elements to assess policies when they scale up beyond the specific experimentation sites, which complements the findings of Vivalt (2020).[8]

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

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

[8]Considerations of the external validity of experimental design have been central to much of the discussions, though it is often focused on individual participants of the policy interventions and experiments, rather than location as a whole (e.g., Snowberg and Yariv, 2018).

2 Institutional background

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

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

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

Conducting policy experimentations before adopting the policies nationwide was later inherited, popularized, and institutionalized by Deng Xiaoping and Chen Yun in the 1980s and 1990s as a core principle guiding the Reform and Opening-up era policy transitions (Heilmann, 2008a; Xie and Xie, 2017). While the earlier policy experimentations during the Communist Revolution and early years of the People's Republic of China typically involve pre-conceived, centrally-imposed model emulation, the policy exper-

imentations during the Reform and Opening-up era are distinguished by their open-endedness in generating novel policy instruments and policy solutions. The "institutional entrepreneurship" released by policy experimentations has long been regarded as a key factor ensuring the stable deepening of China's market reforms (Naughton, 1996).

Primary form of experimentation: *experimentation points* The most pervasive form of policy experimentations in China in the past four decades is through the selection of "experimentation points" (*Shidian*), as noted by Heilmann (2008a,b). Before deciding on whether a new policy should be implemented nationwide, the central government first tries out the policy regionally in a limited number of sites (hence the term "experimentation points") and potentially repeats so for several waves, in order to help evaluate the potential cost and benefit of that policy. Such a gradual approach thus allows effective policy innovations to precede "from point to surface," which could potentially help avoid costly policy mistakes at the national level.

The central government generally announces and introduces the policy experimentations by publishing general guideline documents. Such documents are issued by the ministries and commissions that lead the experimentations, sometimes co-signed by coordinating ministries or the State Council if it involves inter-ministerial coordination. The local government of each experimentation site typically responds to the central government documents by publishing a specific local document of experimentation action plan, laying out logistical and implementational details for the experimentation.

When selecting experimentation points, the central government usually directly assigns certain regions as sites for experiment, but sometimes also solicit local governments who would be willing to voluntarily participate (Zhou, 2013). In either case, both central and local governments need to agree before the policy experiment can take place in a certain region, so there is implicit consent and compliance even for the centrally assigned policy experimentations. Typically, the central government would choose 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 experimentations are 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, which is essentially "creating a new system alongside, or in the interstices of, the existing one" (Naughton, 1996).[9]

[9]The experimental zone aim to explore integrated bundles of policy solutions to economic development, rather than to experimentally evaluate the effectiveness of a specific policy, which is conceptually closer to Sachs (2006). The most noticeable examples for experimental zone would be the Shenzhen Special Economic Zone and Shanghai Pudong Special Economic Zone, which have served as policy laboratories for various

Once a policy experimentation is determined to be successful, certain experimentation points are set as "demonstrative examples," and their experience in implementing the new policy will be actively promoted by the central government to the rest of the country (hence the term, "from point to surface"). Effective policies based on the experimentations would eventually also get formalized in the central government's legal documents and officially become national policies. In contrast, if a policy experimentation fails to generate desirable outcomes, either due to the policy's inherent ineffectiveness, local political economy constraints, high implementation cost, or unexpected public pressure against its implementation, the policy experimentation would quietly stop expanding beyond the initial implementation stage — very few failed policy experimentations are explicitly canceled.

In this paper, we focus primarily on the form of policy experimentations through experimentation points, including those clustered in experimentation zones. Most major reform initiatives in post-Mao China have been tried out by means of experimentation points before they were rolled out to the entire country (if at all); Appendix A.1 describes several other, less common forms of policy experimentations in China. Notable examples of policy experimentations through experimentation points in recent decades include reforms on rural products marketing (from 2000), the reduction of agricultural taxes and levies (2000–2004), reform on land management (2001-2004), and the cooperative health care (since 2003). We describe a particular policy experimentation on the local fiscal empowerment reform in greater detail in Appendix A.2.

3 Data and characteristics of policy experimentations

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

reforms during the Reform-and-Opening era.

3.1 Government documents on policy experimentations

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

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

We identify 663 distinct policy experimentations based on policy themes. Our categorization of policy experimentations is conservative: consecutive experimentations are grouped into the same policy experimentations as long as they concern similar policy aims, even if the specific content of the policies evolves and occasionally even the names of the policies change. Moreover, policy experimentations that are closely related and simultaneous in implementations are combined into one experimentation, even if the central government issues separate documents for each experimentation components.[10] We distinguish different phases of the experimentations by distinct waves of the experimentation roll-out, often marked by specific central government documents.

Coverage of policy experimentations Initiation of experimentation programs from inside the government is by far the most frequent starting point of policy experimentation (Heilmann, 2008b). Government initiated experimentations have corresponding government documents, ensuring our comprehensive coverage on such experimentations. In particular, our data includes extensive coverage on potentially failed experimentations, 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. Specifically, for the ministries that publish documents on their own websites, we independently collect documents from the ministerial websites and

[10]For example, experimentations on corn seed insurance, rice production insurance, professional farmer training, and agricultural tech promotion consulting service are combined together into an overarching experimentation on improving agricultural technology and management.

we find that the government documents collected from the *PKULaw.com* has extensive and comprehensive coverage (see Appendix Table A.1). When we manually examine the limited documents that are published on the ministries' websites but not included in the *PKULaw.com* database, we find that they are secondary documents and do not contain information on policy experimentations.

It is important to note that by relying government documents to describe policy experimentations, the experimentations need to reach a stage of formal endorsement and coordination by the central government in order to be included in our sample.[11] Thus, we do not observe very early stage experimentations initiated by the local governments that never reach to the level of the central government — e.g., early bottom-up policy entrepreneurship led by specific local governments but fail to receive the central government's approval for continuing and expanding the policy. This implies that the set of centrally coordinated policy experimentations that we study is already positively selected in terms of the central government's prior evaluation on the policy's effectiveness. However, such sample selection does *not* mean that policy uncertainty is irrelevant in this context: on average, 45% of the policy experimentations fail to become national policies, even though the central government envisions all of them to become so at the onset of experimentations.

3.2 Time trend of policy experimentations

Figure 1 plots the number of experimentations initiated in each year across the past 4 decades, where we record the first year when a specific policy experimentation started as the year associated with the multi-year roll-out of the experimentation.

We observe similar hump-shaped pattern: the number of policy experimentations initiated by the central government remained relatively low throughout the decades of 1980s and 1990s, averaging less than 10 new experimentations per year across all ministries and commissions. The number of experimentations began to sharply increase towards the end of the 1990s, reaching a peak of 76 new experimentations initiated in 2013 alone. Since 2013, the number of new experimentations started to decline, and nearly halved by the end of our sample period in 2020.

While many factors could contribute to the patterns of the number of policy experimentations initiated over time, at least part of the decline in the number of experimentations 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, fund-

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

ing, and decision rights from the local governments to the upper-level ministerial units, they move away from flat, multi-divisional structures (M-form) to more centralized, unitary structures that benefit from the economies of scale (U-form). We find that following the transition to U-form organization, the vertically managed ministries significantly decrease the amount of policy experimentations they administer. Appendix C presents results using an event study design. Consistent with the theoretical predictions from organization economics (e.g., Chandler, 1962; Williamson, 1975; Qian et al., 2006), the flat, decentralized organizational structure in state bureaucracies provides the flexibility and easiness to coordinate and facilitate policy experimentations.

3.3 Characteristics of policy experimentations

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

First, we extract the experimentation sites and the roll-out schedule of each policy experimentations. Many policy experimentations have more than one wave of roll-out, and we identify 1374 distinct rounds of roll-out across the 663 experimentations. We link each government document to a specific round of experimentation based on its date of issuance, which allows us to observe the time they join a particular policy experimentation and to compare experimentation sites selection patterns across rounds. Table 1, Panel A, presents the total number of policy experimentations initiated during 1980 and 2020, the average number of rounds and experimentation sites involved in each experimentation. On average, each policy experimentation initiated by the central government contains more than 2 rounds in its roll-out and lasts for 2.25 years until either the roll-out stops or the experimentation gets promoted to become a formal, national policy.

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

Third, we observe whether policy experimentations are rolled out to the entire country and become formal national policies. This is marked by specific central government documents concluding the experimentation cycle. Overall, 53.9% of the policy experimentations eventually become national policies and 46.1% fail (see Figure 1, share of successful and failed experimentations indicated by darker and lighter gray shades, respectively). The share of policy experimentations led to national policy roll-out remains remarkably stable over time (see Appendix Figure A.2). The patterns concerning successful and failed policies are not sensitive to the particular definition: for example, we alternative define a policy experimentations to be successful if their roll-out cover at least two-thirds of the whole country's counties, and we find similar patterns throughout (see Appendix Figure A.3).

Fourth, we measure the importance, complexity, and ex-ante uncertainty of each policy experimentations. We capture the degree of importance of policy experimentations by whether they are explicitly mentioned in the central government's *Five Year Plans*, which represent the most important policy blueprints issued by the Chinese government and cover policy agendas considered as highest priority in the upcoming five-year period. 19.2% of the policy experimentations reflect policy themes mentioned in the *Five Year Plans* (see Appendix Figure A.4). We capture the degree of complexity of policy experimentations by the number of ministries and commissions involved in the experimentation, as well as the length of the initiation documents in describing the experimentations. 24.3% of the policy experimentations involve more than 2 ministries and commissions, which we label as complex experimentations (see Appendix Figure A.5). We capture the degree of ex-ante uncertainty of policy experimentations based on whether the central government has laid out detailed national roll-out timeline before the experimentations start. 30.7% of the experimentations feature such timeline (which we label as experimentations on policies with high certainty), and 62.0% of them eventually become national policies. In contrast, among the 69.3% experimentations that do not feature such timeline (which we label as experimentations on policies with high uncertainty), only 36.2% eventually successfully roll out to the entire country (see Appendix Figure A.6).

Fifth, we categorize policy experimentations as either assigned or voluntary, depending on whether the experimentation sites are designated and assigned by the central government directly, or the experimentations invite voluntary participation of the local government. About 40.0% of the experimentations allow (at least partially) for voluntary participation of the local governments (see Appendix Figure A.7).

Finally, we measure a number of auxiliary characteristics of policy experimentations which we incorporate into various parts of the analyses. For example, we identify whether

the central government would provide additional fiscal support for the experimentation sites, and whether the policies at trial would in principle benefit from extra fiscal support. These characteristics would help us evaluate the plausibly strategic fiscal resources allocated by the local governments in order to improve the local outcomes of the experimentations. We also measure how policy innovation and differentiation evolved across time and space, by construct matrices of pairwise textual similarities for all the local policy documents that belong to the same policy experimentations.[12] Such measure would allow us to investigate under what conditions do the local governments exert greater efforts to differentiate their local policy implementations.

4 Is the selection of experimentation sites representative?

Focusing on the policy experimentations that do take place, and we now examine the selection of experimentation sites. As a benchmark, we examine whether the selection of experimentation sites is indeed representative.

From the central government's perspective, a key criterion for experimentation site selection is its representativeness, which determines the quality of knowledge one could extract from a policy experimentation (Zhou, 2013). While the central government likely possesses multiple objectives from conducting the policy experimentations (which we discuss in greater detail below), being able to learn from experimentations in a balanced, representative sample is certainly a central goal. The National Development and Reform Commission, the leading governance body that guides and coordinates national policies, lays out the overall principles of choosing experimentation sites as:

> The balanced distribution of experimentation sites is the most important criteria in choosing these sites. [...] Policy experimentations 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.

4.1 Procedure to test for representativeness

For each policy experimentation, we compare pre-experimentation characteristics between locations where the experimentation is implemented and those that do not par-

[12] Text similarity is calculated using Latent Similarity Analysis, an canonical choice in natural language processing. After removing stop words, we conduct the Tf-idf encoding for each word vector, and then use the first three principal components to compute cosine similarity.

ticipate in the experimentation. As the baseline, we examine the local GDP per capita during the year prior to experimentation roll-out.

We conduct t-tests against the null hypothesis that the pre-experimentation levels of local GDP per capita are indistinguishable among the experimentation sites and non-experimentation sites. This amounts to 663 independent t-tests, one for each policy experimentation.[13] Note that conducting representativeness tests separately for each policy experimentation is conservative: if one were to identify deviations from representativeness with these separate tests, then a pooled test with multiple experimentations will yield more power in detecting unrepresentativeness and reject the null hypothesis. We discuss below the robustness of using a variety of other local characteristics as well as alternative testing methods.

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

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

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

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

The specific context of China's policy experimentations poses two important complications to conducting the representativeness tests. First, policy experimentations are implemented either the provincial level or the county level. The county level experimentations often represent cases where experimentation province(s) are selected by the central government, and the corresponding provincial governments then choose the counties within their jurisdiction to implement the experimentations. Thus, the experimentation sites selections concern distinct administrative samples. We conduct the representativeness tests at the appropriate level for each policy experimentation: for provincial level experimentations the representativeness tests are conducted at the provincial level; and for county level experimentations, the tests are conducted at the county level, stratified based on the experimentation-participating provinces — in other words, counties participating in the experimentations are compared only with other non-experimenting counties

[13] For each policy experimentation's representativeness test, we adjust the respective degree of freedom in the underlying distribution based on the exact share of locations that participate in the experimentation.

within the same province.

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

4.2 Majority of experimentations are positively selected

We plot, in Figure 2, the distribution of the t-statistics on comparing pre-experimentation local GDP per capita between the experimentation and non-experimentation sites. We mark the thresholds of t-statistics where one would reject the null hypothesis of representative experimentation site selection at 90% confidence interval.[14] Table 2 reports the corresponding test statistics (adjusting for the degree of freedoms for each test) and the share of policy experimentations that one can reject the null hypotheses at the 10% level.

We find that the experimentation sites for 80.7% of the policy experimentations are on average richer (in terms of pre-experimentation local GDP per capita) than localities that do not participate in the corresponding experimentations. This share of positive selected implementation of experimentation remains fairly stable over the four decades since 1980 (see Appendix Figure A.8). Even with statistical tests that are fairly conservative, we are able to reject the null hypothesis of representative selection among 50.5% of the experimentations at the 10% level.

The pattern of positive selections of experimentation sites is remarkably robust. First, we observe similar patterns of positive selection when conducting tests using alternative regional pre-experimentation characteristics such as total local GDP, local fiscal expenditure, or local population size (see Appendix Figure A.9, Panels A to C, respectively). Second, the patterns are similar when we separately examine the multi-site experimentations and one-site experimentations: the representative tests are conducted independently for each experimentation among the former, and by pooling consecutive experimentations together among the latter (see Appendix Figure A.10, Panels A and B, respectively). Third, the pattern of positive selection also robust to alternative permutation tests

[14] As discussed above, each of the 663 t-tests has a specific degrees of freedom. We depict visually the average width of the 90% confidence interval (2.36).

among the multi-site experimentations (see Appendix Figure A.11), and to a variety of ways in which we pool and specify random draws of the pooled tests among the one-site experimentations (see Appendix Figure A.12). Finally, the pattern of positive selection is robust if we just examine the subsample of early-round experimentation sites. This effectively holds fixed the number of experimentation sites across policy experimentations when conducting the representative tests, and do not let the level of representativeness to be affected by subsequent rounds experimentation roll-out (see Appendix Figure A.13).

4.3 Unlikely explanations of the observed positive selection

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

Ex-ante policy uncertainty One may speculate that depending on the ex-ante uncertainty that the central government holds toward each policy at trial, the specific objectives of the policy experimentations could differ and thus justify the deviation from representative sample selection. Policy experimentations on policies that are relatively certain that the central government will implement as a formal, national policy (captured by whether the central government specifies a timeline for such roll-out *before* the experimentations start) might not have learning about policy effectiveness as the primary goal. However, when we separately evaluate the degree of representativeness in sites selection among experimentations that are ex-ante certain and those that are ex-ante uncertain (see Table 2, Panel A), we find that the sites selection among ex-ante uncertain policies is in fact substantially higher (average t-statistics $= 2.97$) than that among ex-ante certain policies (average t-statistics $= 2.12$).

Complex experimentations Positive selection of experimentation sites could be justified if richer localities — often represented by better local governance and administrative capacity — may be better at carrying out the demanding trial policies and thus provide more precise signals on the policy effectiveness. Such justification for positive selection could be even stronger for complex experimentations, for example, those that involve coordination and collaborations across multiple ministries and local government bodies. Nonetheless, as shown in Table 2, Panel B, we observe that the sites selection among experimentations that are less complex, involving a single ministry or commission, deviates (sligtly) further from representative than those that are more complex, multi-ministerial experimentations (average t-statistics $= 2.84$ vs. 2.66, respectively).

Eventual scope of policy roll-out Positive selection of experimentation sites could also be justified if the intended geographic scope of the eventual policy is limited to richer localities. While vast majority of the policy experimentations initiated by the central government concerns national policies, there exists different degrees of flexibility in regional targeting across policy domains. Table 2, Panel C, presents the results of the representative tests for experimentations across policy domains. Interestingly, we observe that experimentations on policy domains such as law and jurisdiction that are more likely to be nationally uniform are *more* positively selected (average t-statistics $= 4.53$) than domains such as agriculture that are more flexible to incorporate sub-national targeting (average t-statistics $= 1.98$).

Optimal experimentations Unrepresentative roll-out of experimentations may be justified if the central government as a decision maker carry alternative objectives in addition to learning about the true underlying treatment effects and persuading other agents who might hold different priors. We conduct quantitative exercises to incorporate two specific objectives studied by the recent literature.

First, we examine the incentives of subjective expected utility, in addition to learning and persuasion, following Banerjee et al. (2020). We simulate the optimal experimentation design, parameterizing the model based on China's policy experimentations setup; Appendix D.1 provides details on the simulation procedure. When the central government places heavier weight on its subjective expected utility, deterministic experimentation becomes more justified than randomization. However, even if one places 100% of the weight on decision maker's subjective expected utility, the optimal design of the deterministic experimentation would only induce positive selection with *t-stats* $= 0.813$, which is substantially lower than the positive selection that actually occurs

Second, we examine how the optimal experimentation design would change if the decision maker incorporates considerations over the welfare of the experimentation sites, following Narita (2021). Deviation from full randomization may be justified when the sample size is finite, and there exists sufficiently large heterogeneous treatment effects as well as heterogeneous welfare from receiving the treatment policy.[15] We again simulate the optimal experimentation design, parameterizing the model based on China's policy experimentation setup; Appendix D.2 provides details on the simulation procedure. We find that the central government would have to place almost the entirety of its welfare weights on the locations that were selected as the experimentation sites in the early waves

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

in order to justify the observed degree of positive selection. In other words, the observed level of positive selection could be optimal only if extreme *ex-ante* inequality is inherent to central government's objective function.

4.4 Deviation from representative sample selection due to misaligned political incentives

Could such positive selection occur even if the central government genuinely intends to conduct representative experimentations, as suggested by the National Development and Reform Commission?

Two patterns suggest the potential misalignment between the goal of the central government to learn from balanced, and hence representative experimentations, and the local governments' incentives to positively represent the experimentations.[16] First, we find that experimentations are closer to being representative if the site selection is assigned by the central government directly rather than involving voluntary participation by the local government (see Table 2, Panel D). Second, the deviation from representativeness is not nearly as severe at the province level, as compared to the choices of specific prefectures and counties to be the experimentation sites (see Table 2, Panel E).

We next investigate the political sources of such misalignment more explicitly.

Political patronage Existing literature has documented the prevalent role of patronage in China's political system Fisman and Wang (2015); Fisman et al. (2020). Given the potential political rewards associated with successful policy experimentations, political patronage could also shape the selection of experimentation sites.

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

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

$$y_{mpt} = \alpha \cdot connection_{mpt} + X'_{pt} \cdot \beta + \delta_{mp} + \theta_t + \varepsilon_{mpt}, \tag{3}$$

[16] It has been speculated that local governments may have incentives to deviate from representative sampling, especially by over-sampling richer sites, in order to demonstrate successful experimentation outcomes, which may in turn help with the career advancement of local government officials (Wu, 1995; Huang, 2000).

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

As shown in Table 3, Panel A, when a region becomes connected to a minister, the number of experimentations assigned to that region increases immediately by 28.8%.[17] The effects are almost entirely driven by cases where the central ministry directly assigns the experimentation site, while there is no comparable effect when the experimentation sites are selected via voluntary participation (see Appendix Table A.2). This suggests that the political patronage in experimentation site selection works through top-down favoritism.

Local politicians' career incentives We next examine how the prefectural leaders' career advancement incentive affects their participation in policy experimentations. When local politicians are more likely to be promoted in a certain year, they would have stronger incentives to demonstrate political achievements and to improve their political portfolio, including the participation in important and successful policy experimentations.

Following Wang et al. (2020), we estimate each prefectural city leader's *ex-ante* likelihood of promotion in each year, as a flexible function of their age (relative to retirement), tenure and official rank in the bureaucratic system (capturing the potential for upward mobility); Appendix **??** provides details of the construction of this measure.

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

$$y_{pt} = \alpha \cdot incentive_{pt} + X'_{pt} \cdot \beta + \delta_p + \theta_t + \varepsilon_{pt}, \tag{4}$$

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

[17] In Appendix Figure A.14, we plot the event study estimates around ministers' turnovers. The absence of pre-trend suggests that being connected to a ministry due to central minister turnovers is indeed likely to be orthogonal to the counterfactual trajectories of local governments' experimentation behaviors.

As shown in Table 3, Panel B, when the prefectural leaders have stronger promotion incentives, the corresponding localities engage in significantly more policy experimentations. Importantly, such effects of promotion incentives are almost entirely driven by policy experimentations initiated by M-form ministries (see in Appendix Table A.3). Since the U-form ministries are directly administrated by the central government, the local politicians would have little capacity to influence experimentations initiated by U-form ministries, and more capacity to shape experimentations initiated by M-form ministries directly under the local governments' jurisdiction. This pattern also suggests that our findings are unlikely driven by omitted confounding factors: to the extent that there might be a confounding factor correlated with politician's incentives, it has to be very specifically correlated with only policy experimentations initiated by M-form ministries.

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

In addition to achieving representativeness, the central government could hold other criteria when choosing experimentation sites, and these additional criteria may at times counteract its desire for representativeness. A particular example of such criteria concerns the central government's often overt and overarching objectives to maintain political stability during socioeconomic reforms. Specifically, we examine whether the occurrences of social and political unrest in a particular prefecture affects its chance of being selected as an experimentation site for any policy experimentations in the subsequent period. We find a robust pattern, exploiting within-region across-time variations in unrest occurrences, that prefectures that have experienced social and political unrest in the preceding period are significantly and substantially less likely to become experimentation sites (see Table 3, Panel C). This suggests that unstable local environment could be a veto condition that preclude participation in policy experimentations.We find that the negative relationship between unrest occurrence and selection as experimentation sites are much stronger among policy experimentations with top-down assignments by the central government, than among the ones whose participation is voluntary and bottom-up (see Appendix Table A.4). This suggests that such criteria of avoiding politically unstable locations are primary held by the central government, rather than the potential local participants of the experimentations.

Accounting for observed positive selection Overall, the factors on misalignment between the central and local governments, as well as central government's objectives to maintain political stability could account for nearly 50% of the positive selection in experimentation sites that we observe. We provide several quantitative assessments of these factors in contributing to the experimentation sites selection in Appendix E.

5 Do experimentations induce strategic efforts?

An important component of policy effectiveness is an incentive scheme that encourages sufficient effort from the local governments when they implement the policy. The policy experimentations — perhaps due to its high visibility, high political reward, and explicit monitoring by the central government — may induce additional efforts by the local governments who are especially incentivized to make the policy at trial successful its experimentations stage. In this section, we examine two particular aspects of such experimenter effects: local government's allocation of fiscal resources (Section 5.1) and their efforts to differentiate during experimentations (Section 5.2).

5.1 Allocation of fiscal resources during experimentations

Do the local governments participating in policy experimentations significantly increase fiscal expenditure that may improve local experimentation outcome? To answer this question, we first match each of the policy experimentations to one of the six fiscal expenditure categories: administrative cost, infrastructure, rural production, agriculture/forestry/fishing, culture/education/medicare, and others.

Using a triple-differences strategy, we estimate the following econometric model using county-domain-year level data:

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

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

The results are presented in Table 4, column 1. We observe a substantial increase in domain-specific fiscal expenditure: an additional experimentation increases the local expenditure in the corresponding category by more than 5 percent. The increase in fiscal expenditure is evident especially among experimentations that the central government

does not acknowledge explicit fiscal support in its initial experimentation coordination documents (see Appendix Table A.5).

To examine whether increase in fiscal expenditure related to experimentation crowds out expenditure in other areas, we examine fiscal categories that are least relevant to the policy experimentations. We capture the relevance based on ex-ante correlation across different fiscal categories, and the pairs with the smallest correlations are defined as the least relevant pair. As the local government allocates more fiscal resources to support domains related to policy experimentations, this comes in part at a cost of reduced fiscal support in other policy domains, particularly those that are lest relevant to the policy experimentations (see Table 4, columns 2-3).[18]

5.2 Political incentives and differentiation during experimentations

Next, we examine whether local politicians with stronger career incentives differentiate more in policy implementation details. Differentiation can be a key aspect to signal effort and earn political credits as "model experimentation site."

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

After constructing pairwise text similarity across documents issued by the local governments for a specific policy experimentation p, we then measure each local government i's policy-wide similarity with its peers that have participated in the experimentations prior to or together with the local government i, using the maximum similarity score among these pairs (y_{ip}). We estimate the following econometric model:

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

where $incentive_{ip}$ is the politicians' career incentives as in Section 4.4; X'_{ip} is a set of controls for the politicians (educational attainment and career experience in the central government); λ_i is a full set of location fixed effects; δ_p is a full set of policy experimentation fixed effects; and γ_t is a full set of year fixed effects. Similar to the exercise in Section 4.4,

[18]Such crowd-in of fiscal resources in experimentation-related domain and crowd-out of expenditure in other domains are even larger among policy experimentations that may not require fiscal inputs by the local governments in the first place (see Appendix Table A.6), further suggesting over-expenditure during the experimentations that may not be sustained when the policy at trial become a formal, national one.

we exploit variations in politicians' career incentives due to the timing of the experimentations and their aging relative to retirement.

The results are presented in Table 5. One observes that when local politicians have strong career incentive, they tend to differentiate more from their colleagues in terms of implementation details. While we cannot conclude whether such differentiation is sub-optimal (e.g., if policy solutions that are proven effective has already been tried out by their peers in previous waves of experimentation), the increase in policy implementation differentiation reflects a rise in local politicians' effort on the policy at trial.

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

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

6.1 Central government's policy learning

Policy learning and location-specific shocks When evaluating experimentation outcomes, does the central government exclude location-specific shocks that are orthogonal to the underlying policy effectiveness? In particular, we examine whether local fiscal windfall during the experimentations that may benefit their outcomes affect the central government's policy learning and increase the likelihood that the policy at trial be evaluated successful.

We focus on land revenue (i.e., land conveyance fee) received by the county governments for converting agricultural land for non-agricultural use. This has been one of the most important source of local fiscal revenue windfall in the 2000s (e.g., Han and Kung, 2015). Following the empirical strategy of Chen and Kung (2016), we isolate the exogenous component of such land revenue windfall as a result of two factors (and their interactions): *(i)* the amount of land in a county suitable for commercial and real estate

development as determined by terrain; and *(ii)* exogenous time-varying demand shock of interest rates and regional housing price fluctuations. We evaluate whether the land revenue increase due to these factors unrelated to policy experimentations and policy effectiveness *per se* may affect the chance that policy experimentations become successful and roll out to the entire country.

We estimate the following two-stage-least-squared specification:

$$\begin{aligned} \text{LandRev}_{it} &= \alpha_1 \cdot \text{IV1}_{it} + \alpha_2 \cdot \text{IV2}_{it} + X'_{it}\beta + \delta_i + \gamma_t + \epsilon_{it} \\ y_{it} &= \mu \cdot \widehat{\text{LandRev}}_{it} + X'_{it}\Gamma + \psi_i + \nu_t + \varepsilon_{it}, \end{aligned}$$

where the first instrumental variable (IV1) is interaction term between the geographic constraint on a county's land supply with the temporal variations in the national interest rate, and the second instrumental variable (IV2) is the interaction term between the geographical constraint with the housing prices in China's provincial capitals. y_{it} is an indicator whether the policy experimentation successfully rolls out the to entire country; ψ_i is a full set of county fixed effects; and ν_t is a full set of time fixed effects.[19]

Table 6, column 1, presents the 1st stage results: consistent with Chen and Kung (2016), we find that land suitable for commercial and real estate development, local housing price fluctuation, and their interaction strongly, positively predict the land revenue received by the local government in a specific year. Columns 2-4 then present the 2nd stage results, where we find a consistent and robust positive coefficient of instructed land revenue among experimentation sites on the corresponding policy's national roll-out. This suggests that when policy experimentations are conducted among locations that experience temporal shocks that could improve the policy outcome, the central government does not fully discount such factors and attribute them at least partially to the underlying policy effectiveness, resulting in biases in policy learning.

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

We focus on local politicians' turnovers taking place *after* the beginning of policy ex-

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

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

Specifically, we estimate the following econometric model:

$$y_i = \alpha \cdot \text{turnover}_{ic} + \beta \cdot \text{turnover}_{ic} \times D_{ic} + \gamma_t + \delta_m + \theta_p + \varepsilon_{it},$$

where y_i is the indicator of the policy experimentation i being evaluated as successful and rolls out to the entire country; $turnover_{ic}$ is the indicator of a change in secretary in city c during the experimentation period of policy i; D_{ic} a dummy variable comparing the career incentive of the incumbent at the beginning of policy experimentation and that of his or her immediate successor, following the calculation described in Appendix **??**; γ_t is a full set of year fixed effects; δ_m is a full set of ministry fixed effects; and θ_p is a full set of province fixed effects.

Table 7 presents the results. We observe a consistent pattern that for policy experimentations that are implemented in localities that experience more local political turnover *after* the start of experimentations, the corresponding policy at trial is substantially more likely to be evaluated successful and become national policies. This is especially the case if the local political turnover results in an increase in local politicians' career incentives relative to the out-going politicians. This suggests that when policy experimentations are conducted among locations that experience politician-related shocks that could improve the policy outcome, the central government attribute them at least partially to the underlying policy effectiveness, again resulting in biases in policy learning.

Positively selected experimentations and national roll-out Finally, we document that policy experimentations that are closer to being representative in their implementation are *less* likely to roll out to become nationwide policies. Appendix Table A.7 presents results where we predict the likelihood of policy roll-out based on the underlying t-statistics from the representativeness tests of the corresponding experimentation's sites selection. We observe a robust pattern that deviations from representativeness are associated with *higher* chances of rolling out to the entire country. Intriguingly, this association is much stronger among policy experimentations that are ex-ante uncertain and learning about policy effectiveness from experimentations' outcomes may be first order. While only suggestive, these results are consistent with the interpretation that the central government does not fully take into account that representative experimentations are more likely to

reveal sub-optimal policies, and positively selected policy experimentations may appear to be (artificially) successful.

6.2 National policy outcomes

Information loss due to unrepresentative sample Experimentations reflect the central government's desire to learn about the mapping from policy to outcome. The large heterogeneity in the mapping from policy to outcome would imply information loss due to deviation from representativeness.[20]

We illustrate such information loss first using the context of a specific experimentation on local fiscal empowerment.[21] In order to foster the economic growth across Chinese counties, the central government initiated an experimentation that allows provincial governments to bypass prefectural governments and directly administer the counties within their jurisdictions, effectively providing fiscal empowerment and even fiscal autonomy to the counties participating in the experiment. Between 2003 and 2013, more than 1,100 counties were selected as experimentation sites.

We observe that the experimentation sites are substantially positively selected at the beginning of the experimentation (t-stats of a t-test on pre-experimentation GDP per capita between experimentation and non-experimentation counties are as high as 9.110 during the first two years of the experimentation), and then shifting towards negatively selection (t-stats decrease to -6.156 by 2007) and moving closer to representative selection towards the end of the experimentation (see Appendix Figure A.16).

We find considerable heterogeneity in the effects of such experimentation on local economic development. Using a staggered event study design to estimate the treatment effects on local economic performance among experimentation counties in the early rounds (positively selected) and the later rounds (negatively selected), while controlling for county and year fixed effects, counties that have higher pre-experimentation GDP per capita benefited from the experimentation, while the poorer counties experience worsened subsequent local economic development (see Appendix Figure A.17).[22]

[20]Such information loss would bias the policy learning of the central government and its relevant decision makers even if they are rational and sophisticated. The distortions in policy learning resulted from the unrepresentative experimentations could be even larger, if the key policy makers are not fully sophisticated at processing the existing information and evaluating the outcomes of experimentations.

[21]See Appendix A.2 for details on the institutional background of the local fiscal empowerment experimentation. This experimentation is particularly suitable to examine the degree of information loss due to unrepresentative experimentations, because of the large amount of counties involved in the experimentation as well as a relatively direct measurement of outcomes (namely, local economic performance) of the experimentation.

[22]Such patterns of heterogeneity by pre-experimentation local economic conditions is *not* merely reflect-

In fact, the experimentation, had it been rolled out to the entire country, would generate a net zero effect with both winners and losers (see Appendix Figure A.19 for the distribution of the projected treatment effects of the local fiscal empowerment for each county in China).[23] This case of the local fiscal empowerment experimentation demonstrates that unrepresentativeness of the experimentation, in particular the positive selection at its early stage that disproportionately highlights the positive effects of the experimentation, masks the unequal nature of the fiscal empowerment policy and could in turn bias the central government's policy choices.

Inequality-inducing policies due to similarity with experimentation sites We next examine the overall effects of national policies originated from experimentations: when they roll out to the entire countries, do locations similar to experimentation sites benefit more from the new policy?

For each policy experimentation that eventually leads to a national policy, and locality that do not participate in the experimentation, we calculate a Mohalanobis distance between localities that participate in the experimentation and those that do not (M_{cp}). The distance is calculated based on the vector of pre-experimentation local GDP per capita, local fiscal income, and fiscal expenditure. We then examine, among localities that do not participate in the experimentation, whether national policies differentially lead to faster local economic growth when a specific county is socioeconomically similar to the experimentation sites of the corresponding policy. Specifically, we estimate the following model, exploiting variation from the changing composition of experimentation sites across national policies:

$$Growth_{cp} = \alpha \cdot M_c p + \gamma_c + \eta_p + \epsilon_{cp}, \tag{5}$$

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

ing a general equilibrium effect and in particular, early-mover advantage of the local fiscal empowerment scheme. We observe that even among the poor counties that participated in the experimentation during the early rounds, they experienced a negative treatment effect from the experimentation as well, in magnitudes similar to the poor experimentation sites in later rounds (see Appendix Figure A.18).

[23]We find that provinces with more counties that would have been winners of the fiscal empowerment scheme are more eager to participate in the experimentation. Given the estimated heterogeneous effects of the reform, for each county, we project whether its pre-reform GDP per capita is high enough to experience positive effects — if so, we call it a "winner county" from the perspective of the local fiscal empowerment reform. We use the ratio of winner counties in a given province to capture provincial government's incentives to engage in the experimentation. We find that the ratio of local winner counties in a given province is associated with substantially higher likelihood for the province to be engaged in the experimentation in the first place, and a higher chance that the provincial government emphasize the implementation of this experimentation enthusiastically in its yearly government work report (see Appendix Table A.8).

fixed effects.

Table 8, Panel A, presents the results. One observes that localities that do not participate in the policy experimentations but are more similar to the experimentation sites benefit significantly *more* from the policies when they roll out to the rest of the country. This result is robust to different indices chosen to compute the distance (See in Appendix Table A.9)

These results suggest that policies originated from unrepresentative experimentations differentially benefit some regions over others depending on the sample composition of the experimentation sites. Given that the experimentation sites are overwhelmingly positively selected in terms of local socioeconomic conditions as we shown in Section 4.2, this would generate distributional consequences: positively selected policy experimentations may produce a portfolio of policies that are systematically in favor of certain regions at the expense of others, thus leading to greater inter-regional inequality throughout China.

Indeed, when we trace national policies back to their experimentation stages and the corresponding experimentation sites selection, we observe that a higher number of national policies originated from positive selected experimentations is associated with significantly *higher* subsequent inter-county inequality in economic development (see Appendix Table A.10). Such association between national policies originated from positively selected experimentations and subsequent inter-county inequality is stronger when we focus on the subset of important policies, captured by their appearance in the central government's prominent *Five Year Plans*. On the contrary, this is not the case for national policies that originated from experimentations that are closer to representativeness in their implementation — if anything, those policies are associated with a lower subsequent level of inter-county inequality.

7 Conclusion

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

and national policies originated from the experimentations.

We highlight that policy learning and policy experimentations inevitably take place in complex, political and bureaucratic environment with various status quo constraints and distortions. Existing political and bureaucratic environment could affect the extensive margin of policy learning, namely, how much to experiment in the first place. Such environment could also influence the intensive margin of policy learning — how to structure and implement the experimentations, and how biased the information one may gather from the policy experimentations. Our examination of China's policy experimentations suggest that while experimentations could greatly facilitate policy learning, one needs to pay attention to the manners in which policy experimentations are conducted as more information does not necessarily always result in better decision making.[24] For example, our findings point to a fundamental trade-off that the central government faces: structuring political incentives in order to stimulate politicians' effort that improves policy outcomes, while making sure that such incentives are not exaggerated during the experimentation phase so policy learning remains unbiased. Future works on mechanism design solutions that could improve the efficiency of policy learning could be of great policy relevance and importance.

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

References

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

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

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

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

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

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

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

Besley, Timothy and Stephen Coate, "Centralized versus decentralized provision of local public goods: a political economy approach," *Journal of public economics*, 2003, *87* (12), 2611–2637.

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

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

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

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

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

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

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

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

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

Cheng, Hong, Ruixue Jia, Dandan Li, and Hongbin Li, "The rise of robots in China," *Journal of Economic Perspectives*, 2019, *33* (2), 71–88.

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

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

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

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

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

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

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

Heilmann, Sebastian, "From local experiments to national policy: the origins of China's distinctive policy process," *The China Journal*, 2008, (59), 1–30.

_ , "Policy experimentation in China's economic rise," *Studies in Comparative International Development*, 2008, *43* (1), 1–26.

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

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

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

Jia, Ruixue, Masayuki Kudamatsu, and David Seim, "Political selection in China: The complementary roles of connections and performance," *Journal of the European Economic Association*, 2015, *13* (4), 631–668.

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

Kasy, Maximilian and Anja Sautmann, "Adaptive treatment assignment in experiments for policy choice," *Econometrica*, 2021, *89* (1), 113–132.

Khandelwal, Amit K, Peter K Schott, and Shang-Jin Wei, "Trade liberalization and embedded institutional reform: Evidence from Chinese exporters," *American Economic Review*, 2013, *103* (6), 2169–95.

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

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

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

Maskin, Eric, Yingyi Qian, and Chenggang Xu, "Incentives, information, and organizational form," *The review of economic studies*, 2000, *67* (2), 359–378.

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

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

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

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

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

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

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

Persson, Torsten, Guido Tabellini et al., "Political economics," 2000.

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

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

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

Roberts, Mark J, Daniel Yi Xu, Xiaoyan Fan, and Shengxing Zhang, "The role of firm factors in demand, cost, and export market selection for Chinese footwear producers," *The Review of Economic Studies*, 2018, *85* (4), 2429–2461.

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

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

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

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

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

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

Song, Zheng, Kjetil Storesletten, and Fabrizio Zilibotti, "Growing like china," *American economic review*, 2011, *101* (1), 196–233.

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

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

_ **and Aidan Coville**, "How do policymakers update?," 2019.

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

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

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

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

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

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

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

Figures and tables

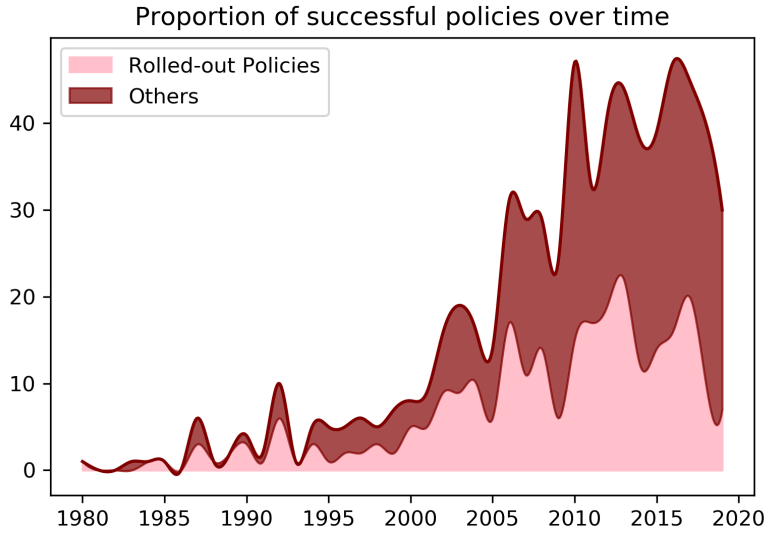


Figure 1: Count of policy experimentations by year. The share of successful experimentations that eventually have rolled out to the entire country is indicated by pink shade; the share of unsuccessful policies that fail to roll out to the entire country is indicated by marron shade.

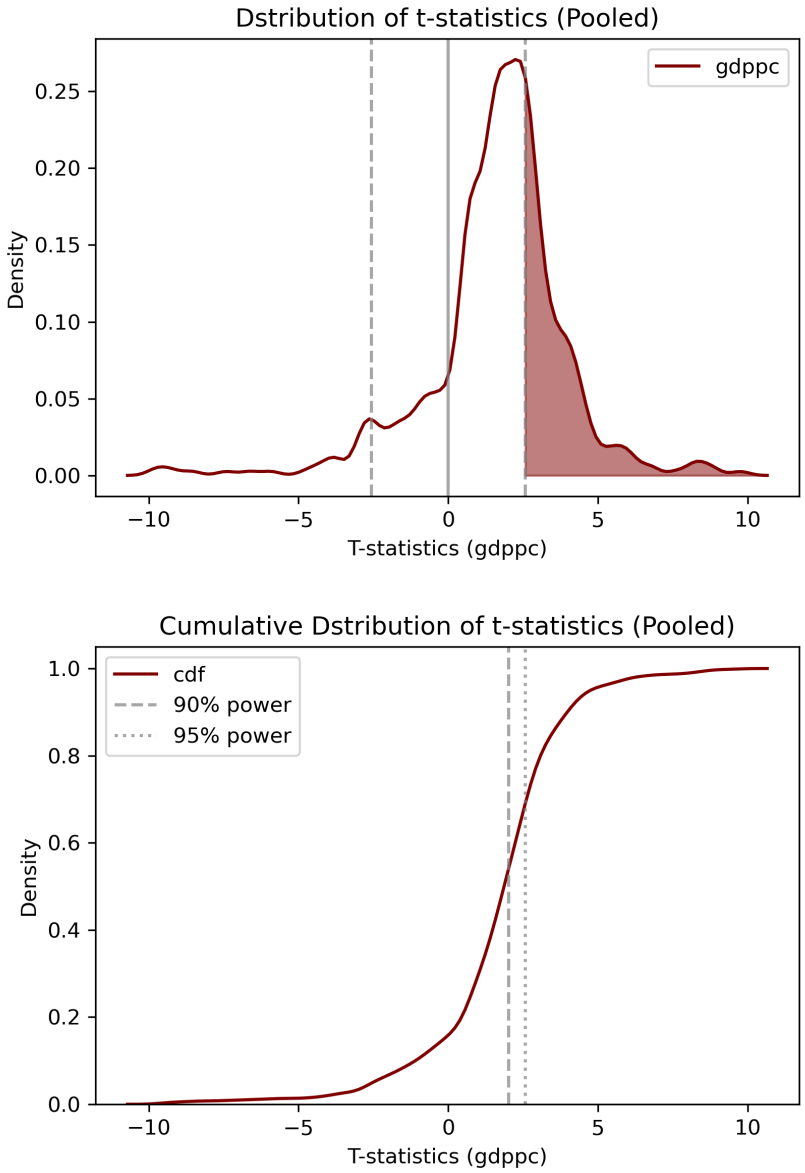


Figure 2: T-stat distribution from the representativeness test, based on GDP per capita. We plot the density in the upper panel and the cumulative distribution function in the lower panel. Within each test we compare the average pre-experimentation GDP per capita between those jurisdictions chosen as experimentation sites, and their peers at the same hierarchical level that are not chosen. The grey vertical lines on both panels shows critical values following a t distribution with degrees of freedom 5, which is close to the average degree of freedom of all tests, at 95% confidence level. Note that those lines are just plotted for visualization purposes. When determining whether a policy substantially deviates from representative site-selection, we compare the t-stats with a critical value adjusted for its respective degree of freedom.

Table 1: Summary statistics of policy experimentations

	# of exp.	# of rounds	# of sites	Roll-out rate
Panel A: full sample				
Full sample	610	2.97	21.00	0.455
Panel B: by policy domains				
Agriculture	56	3.179	10.000	0.321
Market supervision	50	2.660	27.222	0.458
Education	49	2.959	1.500	0.469
Resource, energy & environment	40	2.375	16.435	0.400
Tax & fiscal policy	39	3.205	33.357	0.513
Finance	37	2.432	8.636	0.541
Commerce & trade	29	4.724	4.000	0.414
Population & health	27	3.074	31.556	0.481
Industrial information	25	2.480	20.897	0.400
Domestic affairs	23	2.870	22.440	0.304
Transportation	19	2.053	10.060	0.632
Labor & personnel	11	2.545	46.959	0.636
Law & jurisdiction	9	2.444	89.778	0.667
General purpose	9	4.889	10.308	0.556
Development & reform	3	3.000	12.375	1.000
Media	2	1.000	6.892	0.500

Note: We define a experimentation round if policy documents are issued consecutively within a month, and we report the average number of rounds per experimentation (column 3). "# of sites" (column 4) indicates the number of experimentation sites on average, and "roll-out rate" (column 5) reports the ratio of experimentations that eventually get rolled out nationwide.

Table 2: Tests for representativeness

	T-stats		%
	mean	s.e.	representative
All experimentations	2.74	3.00	47.7
Panel A: by ex-ante certainty			
Certain	2.15	2.85	46.3
Uncertain	2.98	3.04	48.2
Panel B: by complexity			
Single-ministry	2.69	2.82	51.5
Multi-ministry	2.89	3.49	36.3
Panel C: by policy domain			
Law & jurisdiction	4.53	3.43	28.6
Finance	4.34	4.14	45.8
Industrial information	4.21	3.36	36.8
Commerce & trade	4.09	3.21	17.4
Market supervision	3.56	2.47	30.0
Education	2.86	2.79	28.6
Tax & fiscal policy	2.79	3.23	56.3
Population & health	2.26	2.73	38.5
Agriculture	1.98	2.56	65.4
Domestic affairs	1.80	1.66	61.1
Resource, energy & environment	1.69	2.95	77.1
Labor& personnel	1.67	2.01	60.0
State Ministry	1.58	4.79	42.9
Transportation	1.07	1.86	82.4
Media	0.02	-	100.0
Panel D: by sign-up process			
Opt-in	3.01	3.10	43.2
Top-down	2.52	2.90	51.3
Panel E: by administrative level			
Province level	1.13	1.86	71.7
City level	4.35	2.88	28.1
County level	1.57	2.74	56.9

Note: Summary statistics of the t-statistics from representativeness tests — namely, t-tests on pre-experimentation characteristics between experimentation sites and non-experimentation sites. The tests are conducted using pre-experimentation GDP per capita for each administrative unit. Degree of freedoms are adjusted for each policy. Columns 1 and 2 report the mean and standard deviation of the t-statistics, respectively; column 3 reports the percent of policy experimentations that one cannot reject the null hypothesis of representative at 10% confident level. The results are presented visually in Figure 2.

Table 3: Political incentives and policy experimentations

	Engage in experimentations		
	(1)	(2)	(3)
Panel A: Political patronage			
Connected to minister	0.088**	0.062*	0.063*
	(0.035)	(0.036)	(0.037)
Controls	NO	NO	YES
Year FE	NO	YES	YES
Ministry by province FE	YES	YES	YES
Panel B: Career incentive			
Career incentive	1.431*	1.458*	1.353*
	(0.806)	(0.834)	(0.819)
Prefecture controls	NO	NO	YES
Politician controls	NO	YES	YES
Prefecture FE	YES	YES	YES
Time FE	YES	YES	YES
Panel C: Political stability concerns			
# of protests in previous year	-0.004**	-0.002**	-0.003***
	(0.002)	(0.001)	(0.000)
Pre-period controls	NO	NO	YES
Time FE	NO	YES	YES
Prefecture FE	YES	YES	YES

Note: In this table we investigate how various forms of political distortion affect policy experimentations. Specifically, in Panel A, we estimate the increased probability of a province getting assigned a policy experimentation from a given ministry when the current minister used to work in that province. In Panel B, we show how a city leader's career incentive influences the number of experimentations he engages in. In Panel C, we estimate how local protests in the preceding period affect a region's chance of being selected as an experimentation site. In columns 1 to 3, we exploit different specifications in terms of control variables and fixed effects. Standard errors are clustered at province level in Panel A; and prefecture level in Panel B and C. For regression results in other subsamples (M-form/U-form, top-down assignment/voluntary opt-in), please refer to Appendix Tables A.2, A.3, and A.4.

Table 4: Engagement in Experimentations and Local Government Expenditure

	(1) Log (Expenditure)	(2) Log (Expenditure)	(3) Log (Expenditure)
# of experimentations	0.051***	-0.003*	-0.001*
	(0.006)	(0.001)	(0.001)
Expenditure Category	Same	Irrelevant	Residual
Mean of DV	5.090	3.404	8.487
County-Year FE	YES	YES	YES
County-Category FE	YES	YES	YES
Year-Category FE	YES	YES	YES
# of obs.	159416	153458	153458

Note: The standard errors clustered at county level are reported below the estimates. Here we calculated correlation coefficients between every six expenditure categories. For each category, we labeled the least relevant category as "irrelevant", and other categories as "residual".

Table 5: Mayors' career incentive and strategic differentiation

	(1)	(2)	(3)	(4)
	Similarity index			
Career incentive	-0.054**	-0.066**	-0.066**	-0.076**
	(0.027)	(0.032)	(0.033)	(0.038)
Observations	1,206	1,148	1,148	1,148
Policy FE	YES	YES	YES	YES
City FE	NO	YES	YES	YES
Year FE	NO	NO	YES	YES
Politician Controls	NO	NO	NO	YES
Mean of Dep Variable	0.980	0.980	0.980	0.980

Note: Standard errors clustered at policy level are reported below the estimates. We apply Latent Semantic Analysis, a canonical approach from NLP, to measure the text similarity of government documents. The similarity index, taking maximum over all similarity pairs between a document and all others issued by its counterpart administrations on the same policy, aims at measuring how much a local government politician differentiates from his or her colleagues in the policy design during experimentation. Politician controls include his or her level of education and previous central experience.

Table 6: Land revenue windfall and experimentation evaluation

	(1)	(2)	(3)	(4)
	First stage	Second stage		
		# of successful experimentations		
Unsuitability×Interest rate (IV1)	6.004***		-0.122	
	(0.461)		(0.173)	
Unsuitability×Capital city house prices (IV2)	0.257***	0.005		
	(0.027)	(0.007)		
Land revenue (log)		0.048***	0.068***	0.049***
		(0.012)	(0.027)	(0.012)
Observations	10,845	10,845	10,845	10,845
County FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Control variables	YES	YES	YES	YES
Mean of Dep Variable	3.958	0.420	0.420	0.420
Cragg-Donald Wald F statistic		1699.602	61.569	871.581
P value for over-identification test				0.4684

Note: Standard errors clustered at county level are reported below the estimates. County level control variables include the population, tax revenue, and gdp growth rate of counties, all in logarithm. Politician level control variables include his or her age, education, past experience in the prefectural government, previous positions as Youth League party leaders, hometownwise-connection with the prefectural leaders, and years in office. Column (1) reports the first stage results when we include both instrumental variables. Column (2) use the land revenue predicted by our first IV and treats the IV2 as a control variable. Column (3) use the land revenue predicted by our second IV and treats the IV1 as a control variable. Column 4 includes both IV's. We also report the Hansen J statistics for over-identification test, where a rejection casts doubt on the validity of instruments.

Table 7: Political rotation of city officials and experimentation rollout

	National roll-out		
	(1)	(2)	(3)
Politician change	0.045**	0.044***	0.043***
	(0.018)	(0.012)	(0.012)
Change × higher career incentive	0.220***	0.153**	0.151**
	(0.015)	(0.012)	(0.007)
Mean of DV	0.328	0.328	0.328
Province FE	NO	NO	YES
Ministry FE	NO	YES	YES
Exp. end year FE	YES	YES	YES
Obs.	4002	3890	3890

Note: The standard errors clustered at province level are reported below the estimates. At the experiment city level, the career incentive shock is the difference in career incentive between the city official and his immediate successor. Rotate is a dummy variable indicating political turnover during the experimentation, which is defined to be the period between the first round of experimentation and two years after the last round.

Table 8: Similarity with experimentation sites and effects of policy roll-out

	Log growth rate of GDP per capita		
	(1)	(2)	(3)
M-distance	-0.517***	-0.544***	-0.494***
	(0.071)	(0.105)	(0.101)
M-distance * Non-representative		-0.498***	-0.618***
		(0.139)	(0.139)
Mean of DV	0.0806	0.0806	0.0806
County FE	NO	YES	YES
Policy FE	YES	NO	YES
Obs.	77588	77588	77588

Note: Robust standard errors clusters at policies are reported below the estimates. The estimated covariance matrix within the calculation of Mohalanobis distance is fitted by the treated units for each policy respectively.

APPENDIX

A Additional institutional background

A.1 Other forms of policy experimentations

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

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

A.2 County fiscal empowerment reform

In order to demonstrate the potential selection in experimentation site choices and the associated consequences of information loss, we zoom in on the county fiscal empowerment reform as a case study.

Many have argued that when prefectural cities have fiscal control over counties, the lack of fiscal autonomy of rural counties would hinder their economic development (Wang, 2016; Bo, 2020).[1] To address this issue and to foster county economic growth, in 2003, the central government started a large-scale policy experimentation on county fiscal empowerment reform. As illustrated in Appendix Figure A.20, the reform primarily empowers counties by flattening the government hierarchy: before the reform, prefectural cities have fiscal controls over counties, while after the reform, counties can bypass the prefectural government and directly respond to the provincial government.[2]

Within a decade, more than 1,100 counties in China were assigned as the experimentation sites of the reform. The experimentation was rolled out in multiple waves. Based

[1] In the Chinese administrative hierarchy, each province administers several prefectural cities, and each prefectural city administers a number of counties.

[2] The extent of reform also varied across different experimentation points: in some areas, only fiscal power moved down from prefectural cities to the counties (county power expansion), in other cases, all prefectural power was delegated to the counties (province manage counties). For more details on the reform, see Li et al. (2016).

on the central government's document that guides the fiscal empowerment reform, we collect information on the timing at which participating experimentation sites began the fiscal reform.

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

A.3 Government organizational reform

We use the context of China's government organizational reform to understand the organizational environment under which policy experimentations take place in Appendix Section C.

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

The literature on organizational theory distinguishes between two types of organizational structure (Chandler, 1962; Williamson, 1975): multi-divisional form (M-form), which consists of self-contained units in which complementary tasks are grouped together; and unitary form (U-form), which consists of specialized units in which substitutable or similar tasks are grouped together (see Appendix Figure A.21 for an illustration of the distinction between M-form and U-form organizations). While the U-form organizational structure can better take advantage of the economies of scale, the M-form structure provides more flexibility for experimentation. Under the M-form, local managers are able to ensure attribute matching across multiple dimensions, which makes it easier to carry out local experimentation. In contrast, under the U-form, inter-organizational coordination is needed to achieve attribute matching, which complicates potential experimentation (Qian et al., 2006).

The vertical management reforms took place in a staggered fashion over an extended period of more than two decades. See Appendix Table A.11 for a list of the ministries that underwent the vertical management reforms and the years at which they took place.

B Auxiliary data sources

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

B.1 Biographical information of politicians

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

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

Specifically, we estimate each prefectural city leader's *ex ante* likelihood of promotion in each year, as a flexible function of his age when starting the term/position, position and official rank in the bureaucratic system. Our data documents observations across 4980 terms of office, in 333 prefectural cities in China from 1985 to 2017. At the politician level, we document his birth age, educational background, current hierarchical level in the government, previous work experience and promotion status after the term.

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

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

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

Note that t here stands for term of office. The observational level is prefecture by term, so the career incentive index we constructed will be a fixed value throughout a given term of office. Appendix Table A.12 shows the estimated coefficients in the first stage. The first 2 columns shows estimates by LPM and column 3 and 4 shows estimates by Probit. The sign and magnitude of the estimated coefficients are consistent with Table 2 from Wang et al. (2020).

B.2 Government organizational structure

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

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

B.3 Local socioeconomic conditions

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

B.4 Local fiscal expenditure

Fiscal data 93-06 from Tsinghua University

B.5 Land revenue of the local government

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

We match land revenue data (based on Fiscal Statistical Compendium for All Prefectures and Counties, from which data is available for the period 1999–2006, and the website of the Land Transaction Monitoring System, http://www.landchina.com, for 2007-2008 data) with geographic elevation data from United States Geographic Service (USGS) Digital Elevation Model (DEM) at 90-meter resolution, which allows us to estimate the percentage of land unsuitable for real estate development. Moreover, we match the land revenue data with the housing price data from the *Statistical Yearbook of Regional Economics* (2000-2009), which proxies for land demand. We used the interaction of both as an instrumental variable for the land revenue income of local government. The construction of such instrumental variable follows essentially that of Chen and Kung (2016).

B.6 Five Year Plans

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

B.7 Local political and social unrest

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

C Organizational structure and experimentation tendency

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

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

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

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

We formally examine whether the M-form organizations in government bureaucracy are better at facilitating policy experimentations, and U-form organizations are relatively worse at coordinating and initiating such experimentations. In particular, we identify

the impact of a M-form to U-form transition on the number of policy experimentations initiated by the ministry or commission. Following an event study design, we estimate the following specification:

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

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

Appendix Figure A.22 plots the non-parametrically estimated D^k_{mt} coefficients. Consistent with the theoretical predictions, following the transition to U-form, we find that the vertically managed ministries significantly decrease the amount of policy experimentation they administer. The decrease is substantial in magnitude, representing a 59.4% reduction in the number of policy experimentation initiated over the first three years after the organization restructuring, relative to the average level just prior to the U-form transition. Suggesting a causal interpretation, we do not find any noticeable pre-trend leading up to the U-form transition; in other words, there does not appear to be strategic timing of the U-form transition targeting ministries or departments on particular trajectories in terms of policy experimentations they initiated, neither are there substantial preemptive experimentations just prior to the transition away from M-form organization.

Taken together, the results presented above indicate that the flat, decentralized organizational structure provides the flexibility and relative easiness to coordinate, which in turn facilitates policy experimentation. At least part of the decline in the number of experimentations in the recent decade that we observe is due to a shift away from the flat, multi-division organizations of the state ministries to a more centralized structure that benefit from the economies of scale, which may be an inevitable outcome as the development reaches a relatively high and mature level. A simple back of the envelop calculation suggests that one could attribute a reduction of 5 policy experimentations per year to the shifts of ministries to U-form. Though importantly, such a shift to U-form organizations that benefit from the economies of scale may push against the *increasing* need for policy experimentations, as reforms and the policy space become more complex and uncertain with the social and economic development.

D Optimal experimental design simulations

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

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

Overview First, we examine the incentives of subjective expected utility, in addition to learning and persuasion. Following Banerjee et al. (2020), we simulate the optimal experimentation design, parameterizing the model based on the experimentation setup and estimated heterogeneous treatment effects from Section **??**. As predicted by Banerjee et al. (2020), when the decision maker (central government) places heavier weight on its subjective expected utility, deterministic experimentation becomes more justified than randomization. However, even if we place 100% of the weight on decision maker's subjective expected utility, the optimal design of the deterministic experimentation would only induce positive selection with *t-stats* $= 0.813$, which is substantially lower than the positive selection that actually occurs. We further extend the simulation exercise among all policy experimentations, and we conservatively assume that the heterogeneous treatment effect would be twice as large as one observes in the case of fiscal empowerment reform; we obtain qualitatively similar conclusions. Under reasonable assumptions, motivations to maximize subjective expected utility alone is *not* able to justify the level of deviation from representativeness in experimentation site selection that we observe.

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

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

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

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

We simulate the optimal experimental design for both the fiscal empowerment reform and a general experiment with the following procedure:

1. For the fiscal empowerment reform, we first compute the vector of treatment effects

p for each county that receives treatment, using a difference-in-difference specification with controls for pre-experiment GDP and province fixed effects. There are 597 counties that receive treatment during the first waves. Using these treatment effects, we then impute treatment effects for the non-treated group based on the covariates GDP and province. The total sample consists of 2,010 counties, and the mean treatment effect is an increase in GDP per capita of 2.09% over the pre-period quantity (s.d. = 23.09%).

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

(a) For our general simulations, we aim to capture a set of plausible treatment effects that a policy might create. Thus, we generate treatment effects following the data generation process:

$$\beta_c = GDP_c + \zeta_c * \sigma \qquad \zeta_c \sim U[-GDP_{max}, GDP_{max}]$$

where c indexes a county, β is the treatment effect, GDP is the county's GDP, GDP_{max} the largest observed GDP, and σ a parameter that determines the variance of the treatment effect. Thus, the treatment effect consists of a baseline source of heterogeneity plus a noise term.[3] Note that the magnitude of the mean treatment effect size only matters in context of the variance, since the DM is always choosing between a control (treatment effect $= 0$ for all counties) and the policy. Thus, varying σ is sufficient to give us the range of potential scenarios that a DM faces. We display results for $\sigma \in \{\frac{1}{1000}, \frac{1}{100}, \frac{1}{10}, 1, 10, 100, 1000\}$.

(b) For the simulations at the prefecture and province level, we compute treatment effects using prefecture or province level data instead of the county level data. The same applies for the remaining steps.

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

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

where p_s indexes a particular sub-prior, c indexes a county, β is the true treatment effect, $\bar{\beta}$ the mean treatment effect, and β_{max} the largest observed treatment effect. Hence, the sub-prior can be broken into three terms: the true treatment effect β_c, an idiosyncratic bias on the effect of the treatment for each prior γ_p, and random noise $\eta_{p,c}$. Hence, the expected value of each sub-prior's treatment effect is the true

[3] We use GDP as a baseline source of heterogeneity for treatment effects given that it plausibly affects many different policy outcomes. However, this could be substituted for other observables without theoretically affecting the qualitative results.

treatment effect.[4] We construct 1,000 priors to form H and run the simulation with the DM holding each of these priors as their own (h_0) with the other priors treated as hostile.

3. Then, we construct the space of potential solutions to the DP. A solution to the DP consists of an experimental design ϵ and an allocation rule α. Each experimental design is a randomly selected draw of half of the counties to be treated in the experiment.[5] 1,000 of these experimental assignments are generated in the simulation. The allocation rules take the form

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

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

   (a) Given that there are $\binom{2010}{1005}$ potential ways to select an experimental assignment, it is likely that the most optimal assignment for any given prior will not be selected by the simulation—which may matter if the optimum has a vastly more extreme t-statistic compared to the best randomly generated assignment. To remedy this, we add to the 1000 randomly generated assignments a further 1000 optimal assignments, one for each prior. We formulate finding the optimal experimental assignment for a given prior as a Mixed Integer Linear Programming (MILP) problem:

   Let P be the number of sub-priors contained within a prior, n be the number of counties, X be a vector of dummies x_c where $x_c = 1$ iff county c is treated, and Y_s a vector of expected treatment effects for sub-prior s. Then, the optimal assignment is the solution to the MILP problem with the following objective function and constraints:

$$\begin{aligned} \max \quad & e_1 + \ldots + e_P \\ & s.t. \\ & \forall s \in [P], e_s = 1^T Y_s \delta_s + 0(1 - \delta_s) \\ & \forall s \in [P], X^T Y_s \geq -M(1 - \delta_s) \\ & \forall s \in [P], X^T Y_s \leq M\delta_s \\ & \sum_i x_i = \frac{n}{2} \\ & \forall i \in [n], x_i \in \{0, 1\} \end{aligned}$$

   where we use big-M to represent an arbitrarily large constant. Note that e_s is the expected value of assignment X under sub-prior s, and $\delta_s = 1$ iff $X^T Y_s \geq 0$.

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

[5] In the case of the fiscal policy experiment, we instead draw 597 and 1,168 counties, corresponding to the number of counties treated in the initial stage and after the full experiment.

We use a MILP solver to find the actual solution to the problem, generating an optimal design for a given prior.

Since each of these 1,000 optimal designs also has 5 potential allocation rules, this forms 5,000 additional potential solutions to the DP, for a total of 10,000.

4. Once the priors and potential DP solutions have been constructed, we proceed to maximize the DP by finding the optimal solution for each prior $h \in H$. We solve four versions of the DP for each prior, corresponding to $\lambda \in \{0.25, 0.50, 0.75, 1.00\}$. For each of these (deterministic experimental design) solutions, we then compare its expected value to the expected value under the RCT experimental design (where the set of sampled experimental designs is taken as representative of the total), and select whichever is higher as the optimal solution.

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

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

This process yields the following results:

- $\lambda = 1.00$: mean t = 0.746, RCT optimal 14.3% of the time
- $\lambda = 0.75$: mean t = 0.743, RCT optimal 33.7% of the time
- $\lambda = 0.50$: mean t = 0.728, RCT optimal 43.7% of the time
- $\lambda = 0.25$: mean t = 0.748, RCT optimal 73.4% of the time

D.2 Simulations with welfare considerations following Narita (2021)

Overview Second, we examine how the optimal experimentation design would change if the decision maker incorporates considerations over the welfare of the experimentation sites. Narita (2021) demonstrates that deviation from full randomization may be justified when the sample size is finite, and there exists sufficiently large heterogeneous treatment effects as well as heterogeneous welfare from receiving the treatment policy (either captured as experimentation subjects' willingness to pay, or benevolent social planners' welfare weights across subjects). We again simulate the optimal experimentation design, parameterizing the model based on the experimentation setup and estimated heterogeneous treatment effects from Section 6.2. We find that the central government would have to place almost the entirety of its welfare weights on the locations that were selected as the experimentation sites in the early waves in order to justify the observed degree of positive selection, suggesting that the observed positive selection of experimentation sites could be optimal only if extreme *ex-ante* inequality is inherent to central government's objective function.

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

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

2. Subjects maximize utility, satisfying

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

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

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

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

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

Once the set of optimal $(p^*_{c,t})$ are found, we use them to draw 1,000 experimental designs and compute the average t-statistic for this set of WTPs. We collect these average t-statistics over a set of 1,000 sets of potential WTPs, yielding an average optimal t-statistic of 1.188.

E Accounting for positive selection of experimentation sites

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

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

Appendix Table A.13 shows the marginal effect of Log GDP per capita on the probability of being chosen as an experiment site. Positive selection bias is observed across columns. In columns 2 and 4, when those political distortions are controlled, the regression coefficients reduces to only half the amount without controls.

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

Appendix Figure A.23, Panel B, shows the distribution of t statistics of the representative test, as described in Section 4.1, when we assert a non-stochastic version of treatment assignment mechanism. In this setting, those prefectural units with the top k propensity score get chosen as experimentation spots, where k corresponds to the number of sites chosen for each policy at status quo. Compared with our baseline specification shown in Appendix Figure A.23, Panel A, we observe positive selection bias of even greater magnitude. This is consistent with the strict nature of the non-stochastic assignment of policy experimentations.

For a milder version, we plot the distribution of t statistics of the representative test, when we assign experimentation sites stochastically according to their fitted propensity scores within each policy. For simplicity, we assume the sampling proceed is i.i.d., and the number of experimentation sites remains the same as that chosen at status quo. We conduct 1,000 simulations and plotted the pooled results in Appendix Figure A.23, Panel C. This specification is most similar, in general ideas, to the regression presented in Table A.13, confirming the idea that all distortion factors we identified explain almost half of the selection bias of policy experimentation.

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

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

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

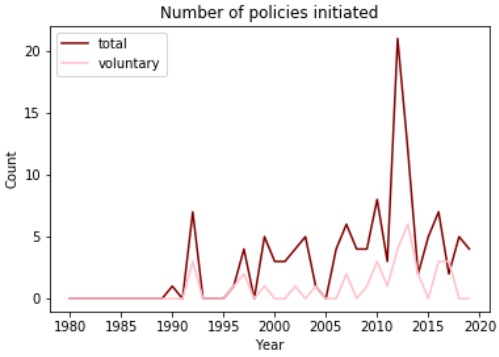
$$w_{ij} = c_{ij} \ln\left(\frac{N}{n_i}\right),$$

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

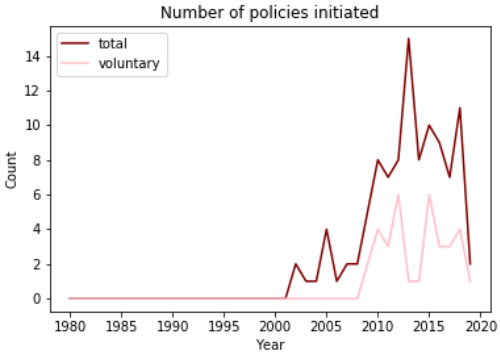
$$\begin{aligned} U_D \Sigma_D V_D^\top &= \arg\min \|M - M_d\|_F^2 \\ \text{s.t.} \quad &\text{rank}(M_d) = D \end{aligned}$$

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

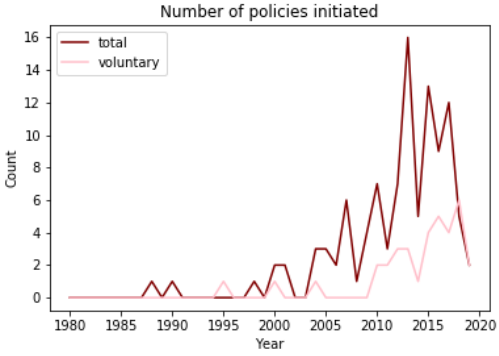
G Additional figures and tables



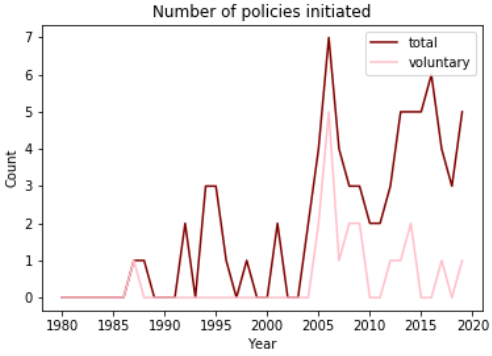
Education



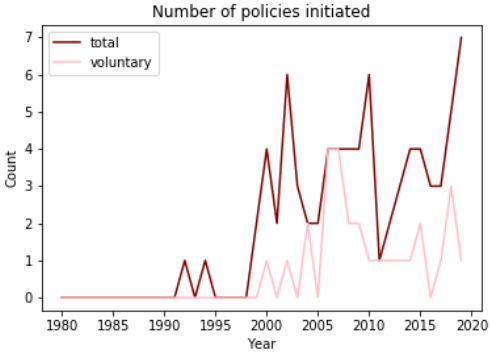
Agriculture



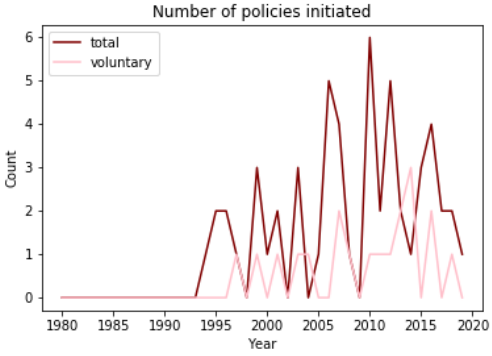
Market supervision



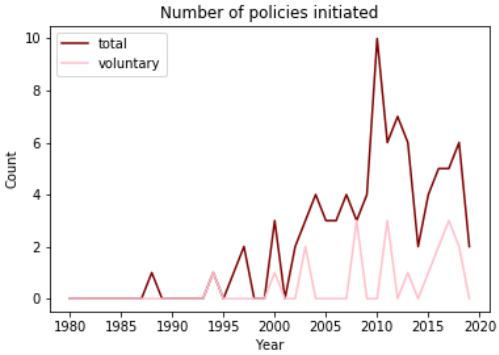
Finance & economics



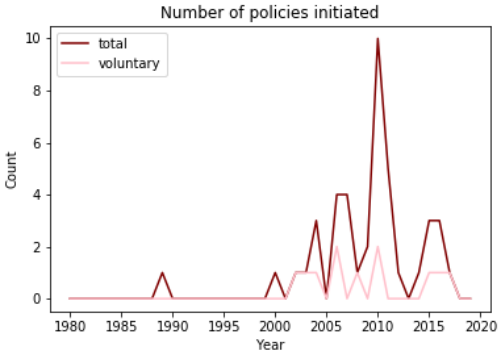
Natural resources & environment



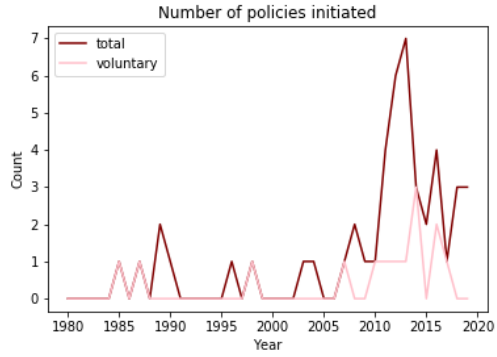
Business & commerce



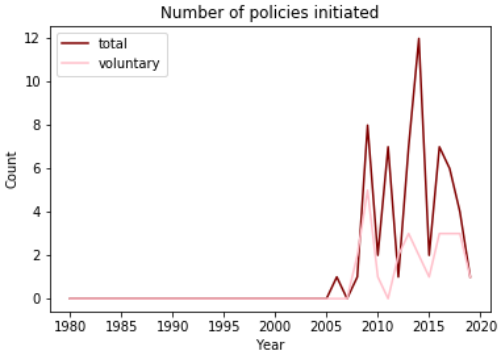
Government finance & taxation



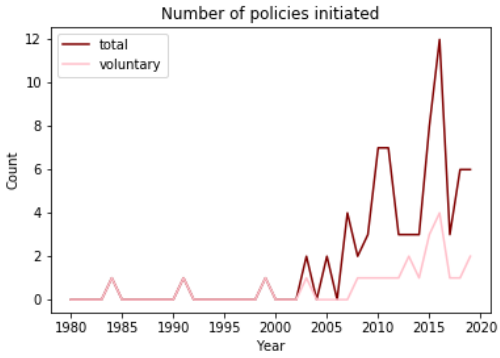
Population & health



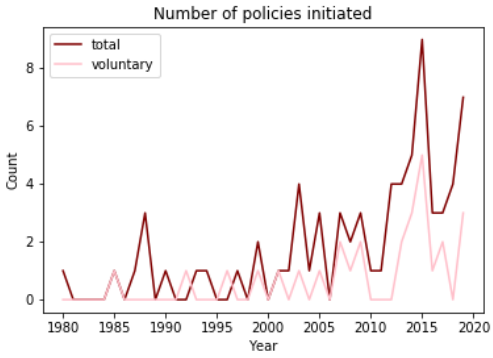
Domestic affairs



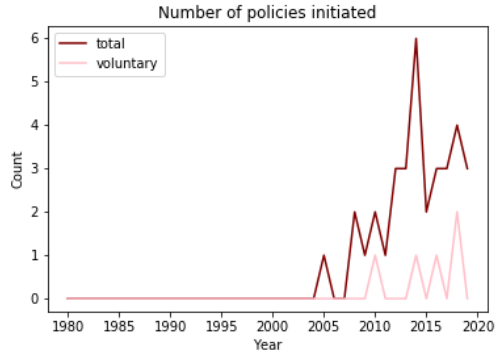
Industrial information



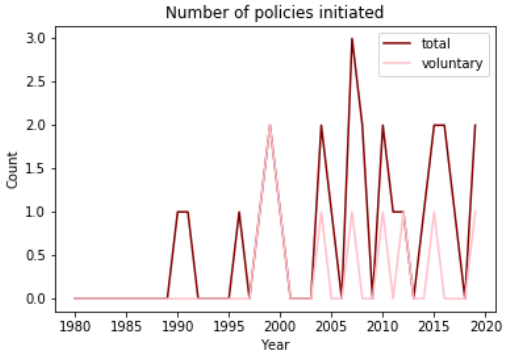
Development & reform



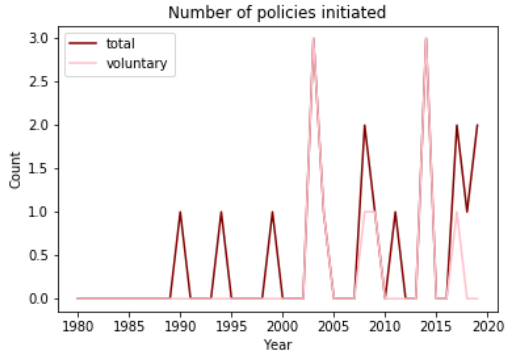
General purpose



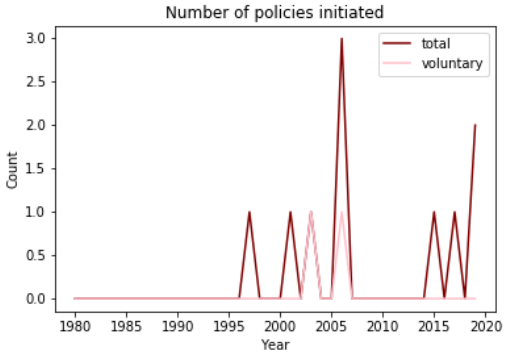
Transportation



Labor & personnel



Judiciary & supervision



Media

Figure A.1: Count of policy experimentation, by ministerial function.



Figure A.2: Time trend of successful policy shares, 2000-2020. Prior to 2000, we have less than 10 experimentations per year on the denominator. Trends are very similar with greater fluctuation.

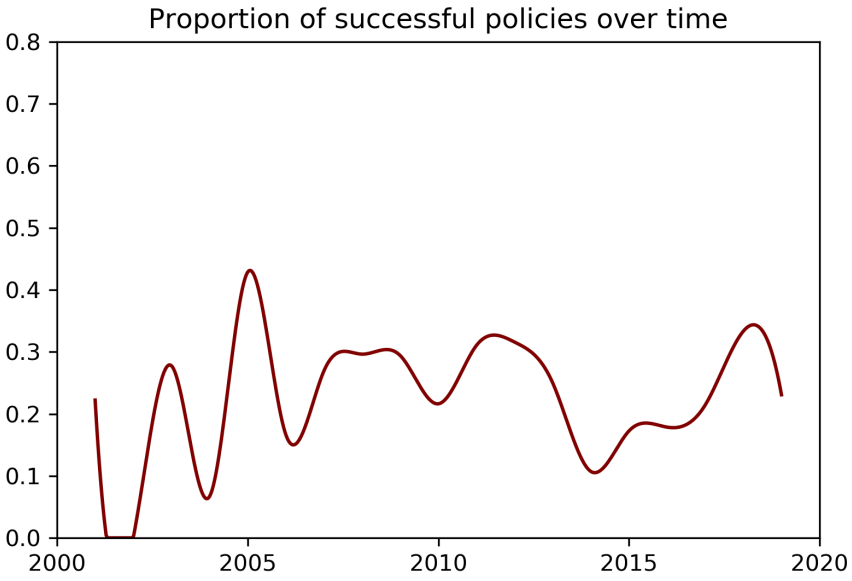


Figure A.3: Time trend of successful policy shares, 2000-2020. Prior to 2000, we have less than 10 experimentations per year on the denominator. Trends are very similar with greater fluctuation.

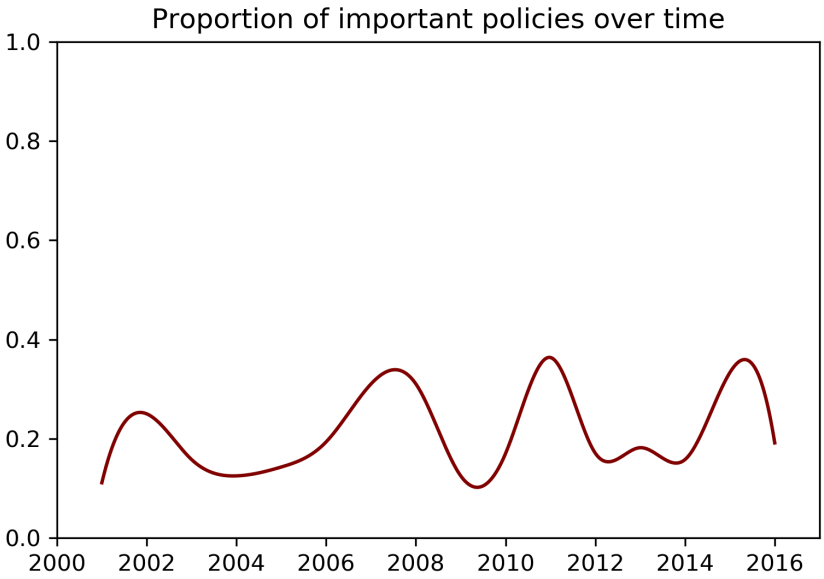


Figure A.4: Time trend of important policy shares, 2000-2017. A policy is labeled important if it is mentioned in the Five Year Plans. Those plans are both retrospective and introspective. We dropped some of the most recent years observing the most recent Five Year Plan is issued in 2016.

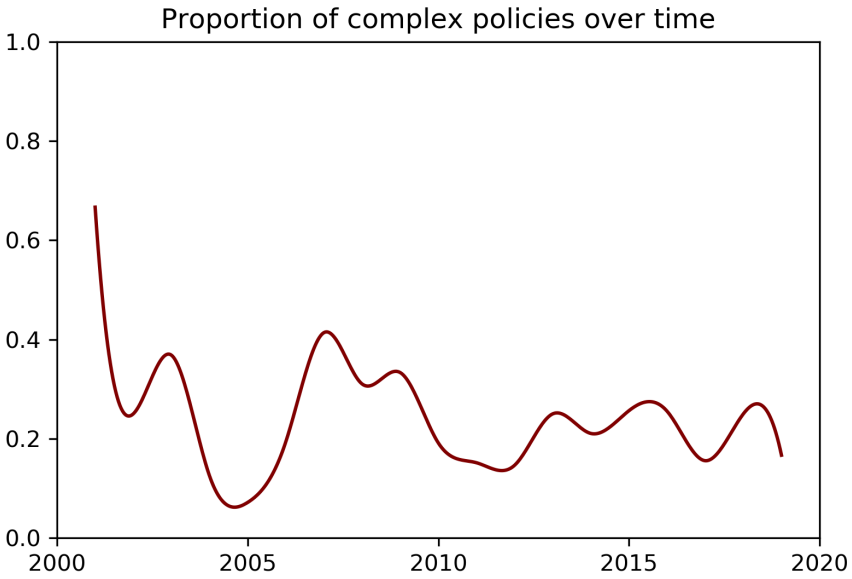


Figure A.5: Time trend of complex policy shares, 2000-2020. Complex policies are defined as policies that require multi-department cooperation.

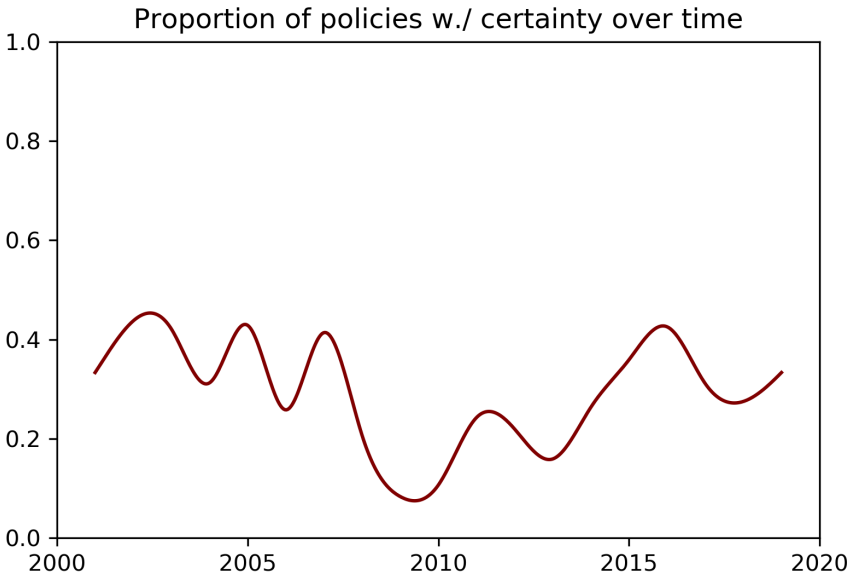


Figure A.6: Time trend of certain policy shares, 2000-2020. A policy is labeled with certainty if detailed timelines of rollout is delineated at the first document issued with regards to the experimentation.

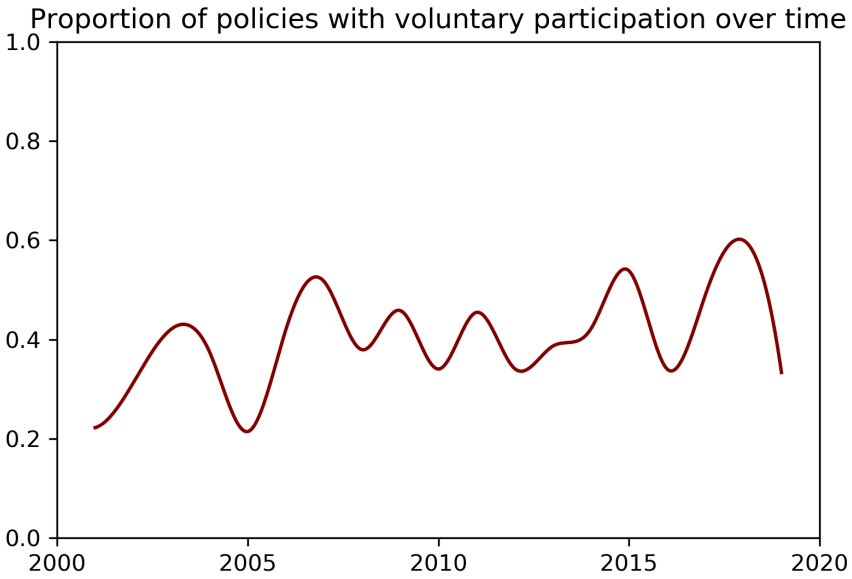


Figure A.7: Time trend of voluntary policy shares, 2000-2020. Voluntary policies typically have a sign-up process. The central government do not directly assign experiment sites at the initial document for experimentation.

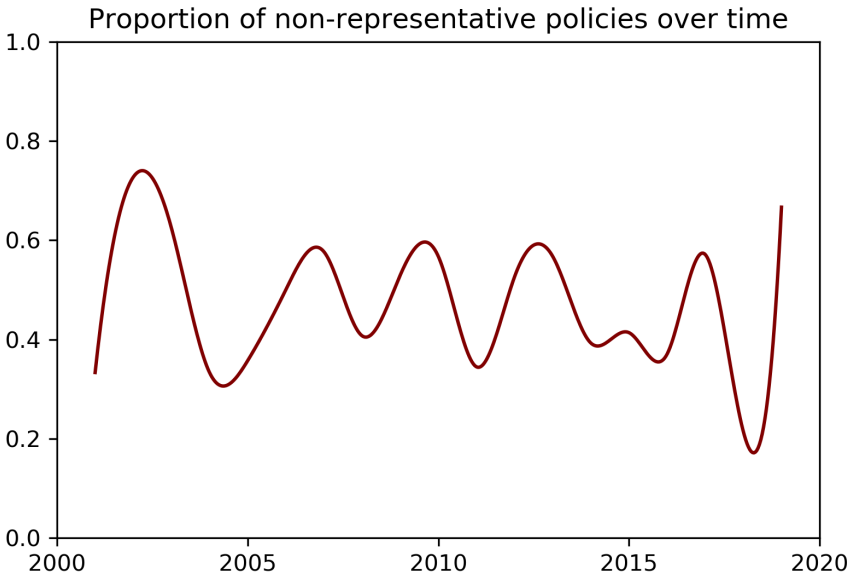


Figure A.8: Time trend of non-representative policy shares, 2000-2020.

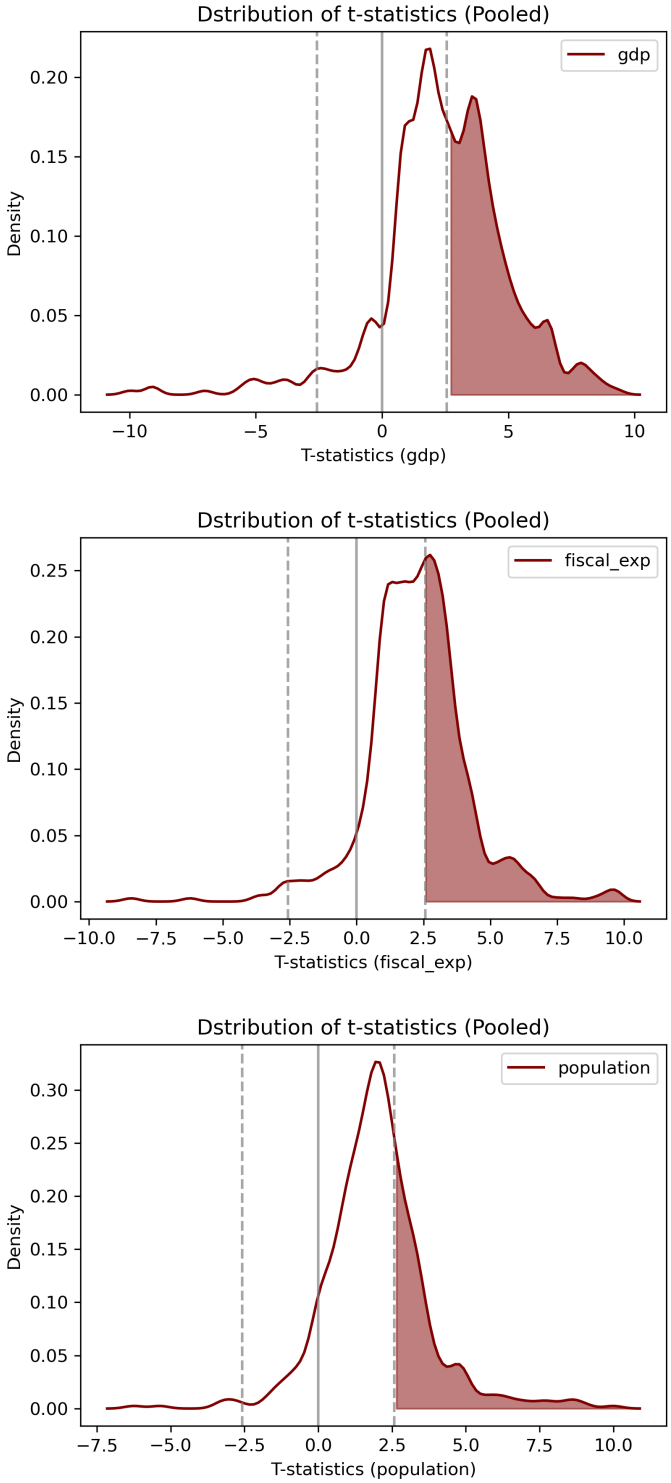


Figure A.9: Representativeness tests' t-statistics distribution, using other features including GDP, fiscal expense and population, with the same test procedures as Figure 2.

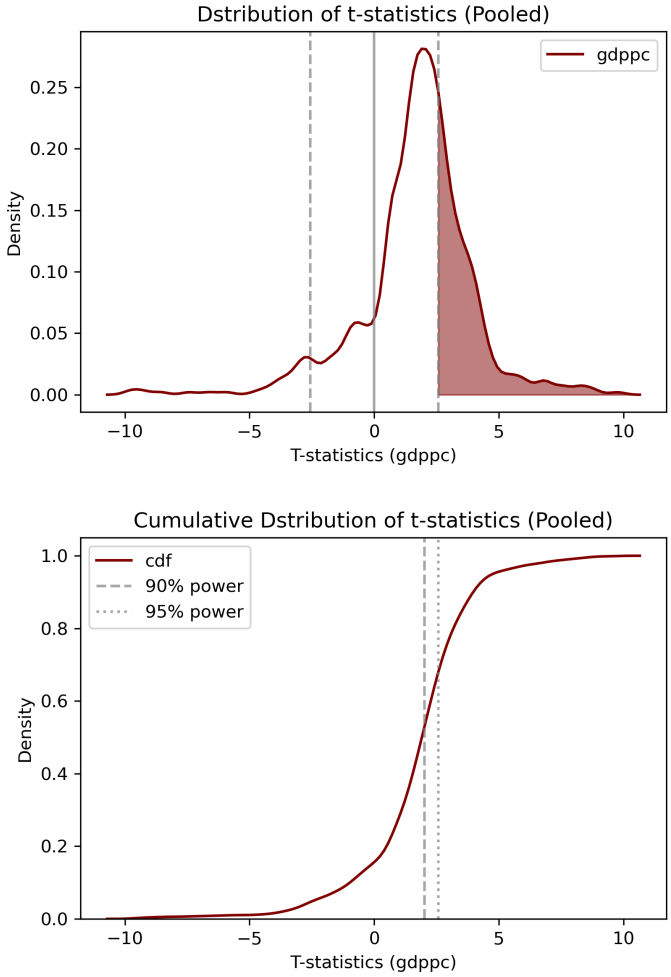


Figure A.10: Representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 2. One site policies are pooled chronologically by clusters n=5.

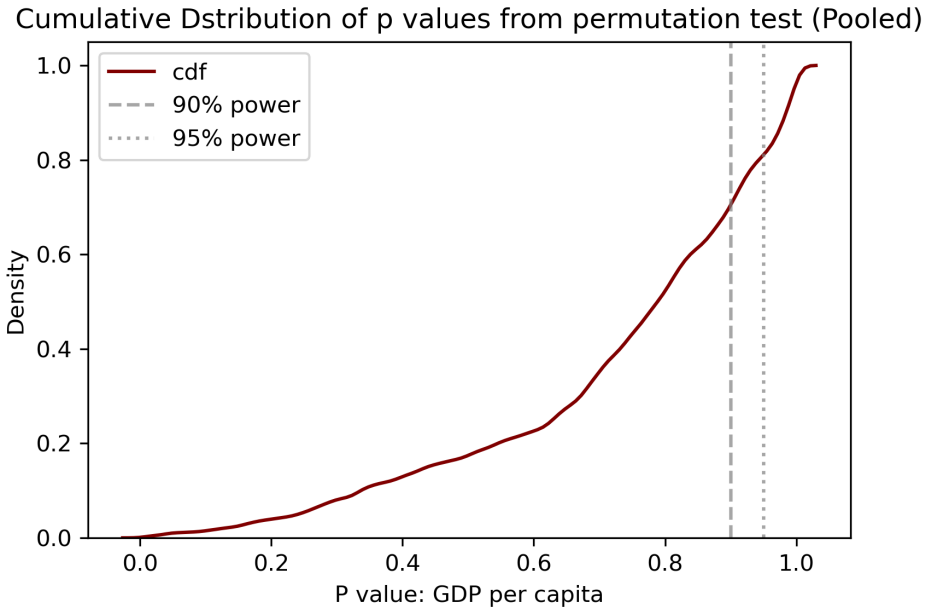


Figure A.11: Cumulative distribution of p values from permutation tests in representativeness tests. Each realized student-t statistic is compared with 5,000 permuted t values to calculate the p statistic. In small samples, permutation tests are more conservative than standard t tests.

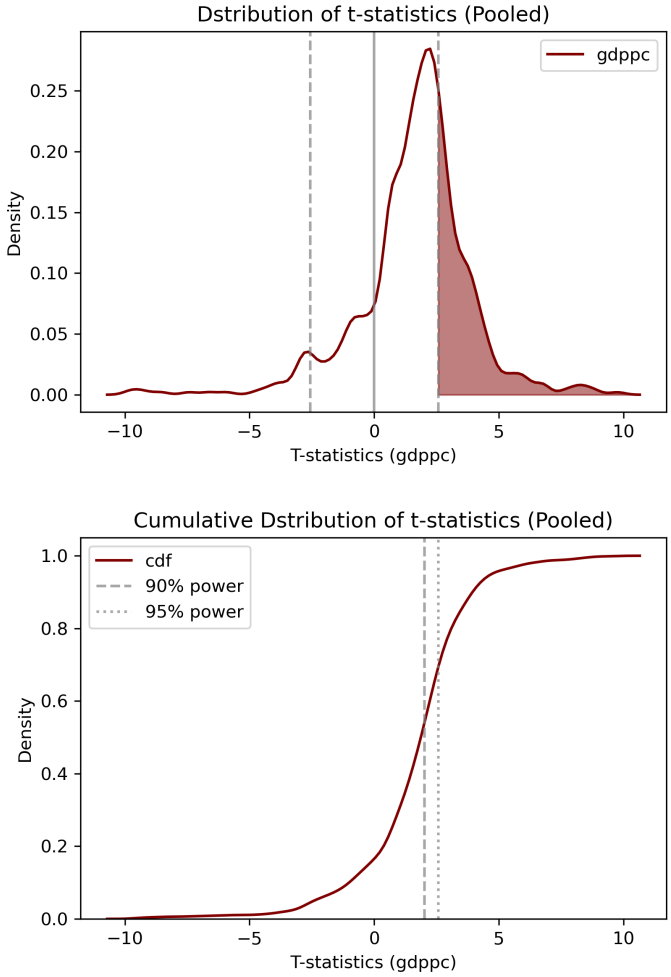


Figure A.12: Representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 2. One site policies are pooled by bootstrapping with replacement for 166 times, without specifying the existence of any certain policy, and concatenated to the multi-site sample.

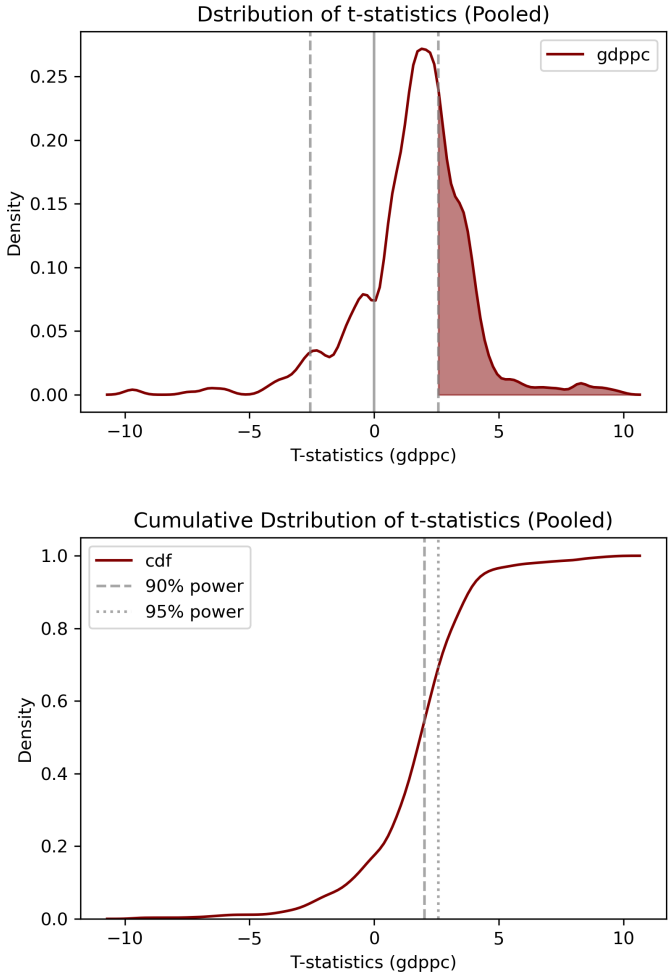


Figure A.13: Representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 2. Only policies in early rounds are considered.

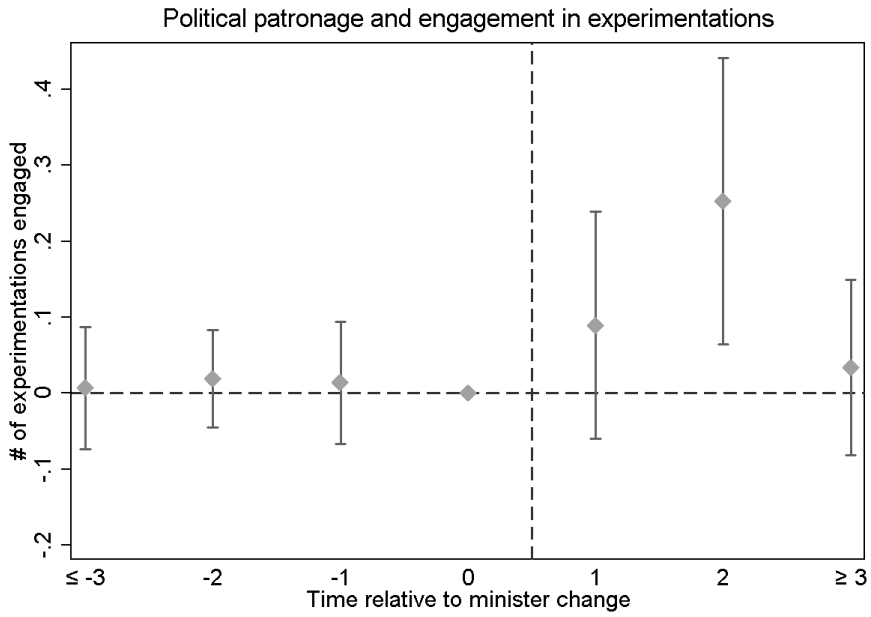


Figure A.14: Probability of being selected as site for experimentation after being related to a ministry due to central minister turnovers.

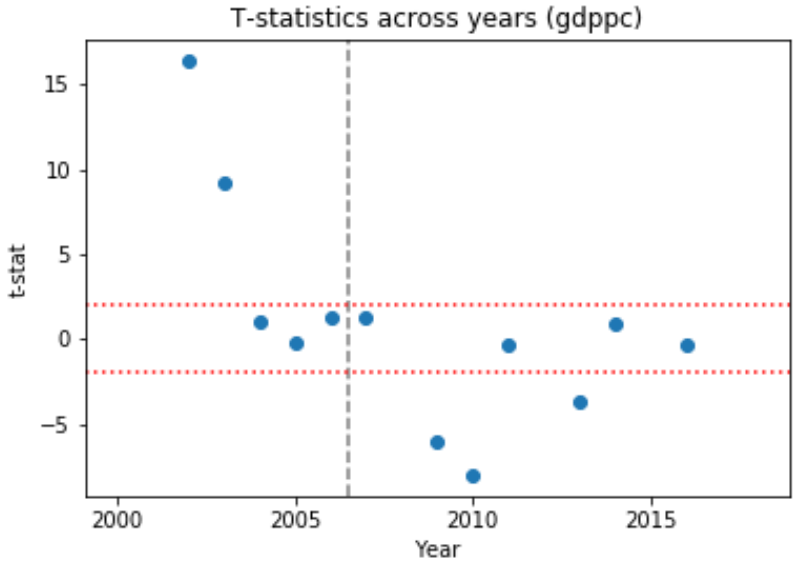


Figure A.15: Local fiscal reform — representativeness test, without province FE. Same test procedure as Figure 2. No provincial clusters are taken into account. Compared with the last figure, the underlying assumption of this test is that the choice of treated units is decided entirely by the central government, whereas in Figure A.16 we assume that the decision is made at provincial government level.

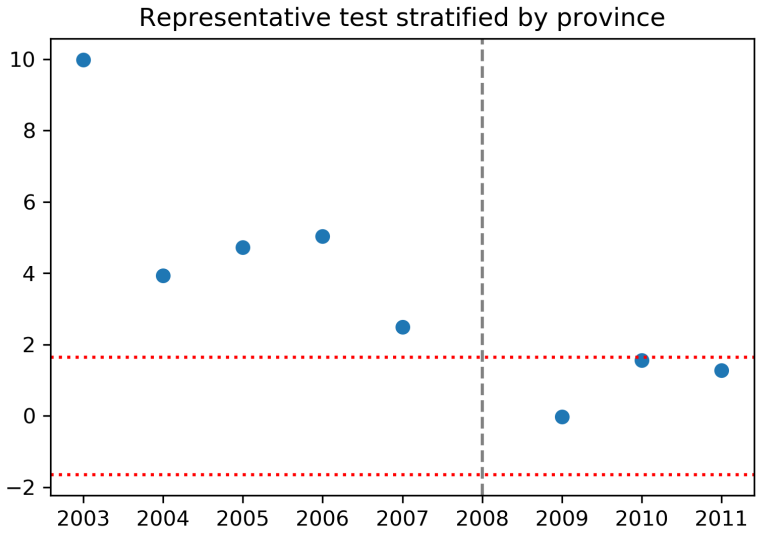


Figure A.16: Local fiscal reform — representativeness test. We conduct stratified Fisher randomization tests with student-t statistics and provincial strata. Within each province, we view counties that engage in the experimentation for the first time as units of the treatment group, the rest as control. Provincial level t-stats are weighted and standard errors are estimated based on Miratrix et al. (2013). The red horizontal lines indicate the asymptotic 90% confidence intervals within which representative assignment of experimentation sites cannot be rejected.

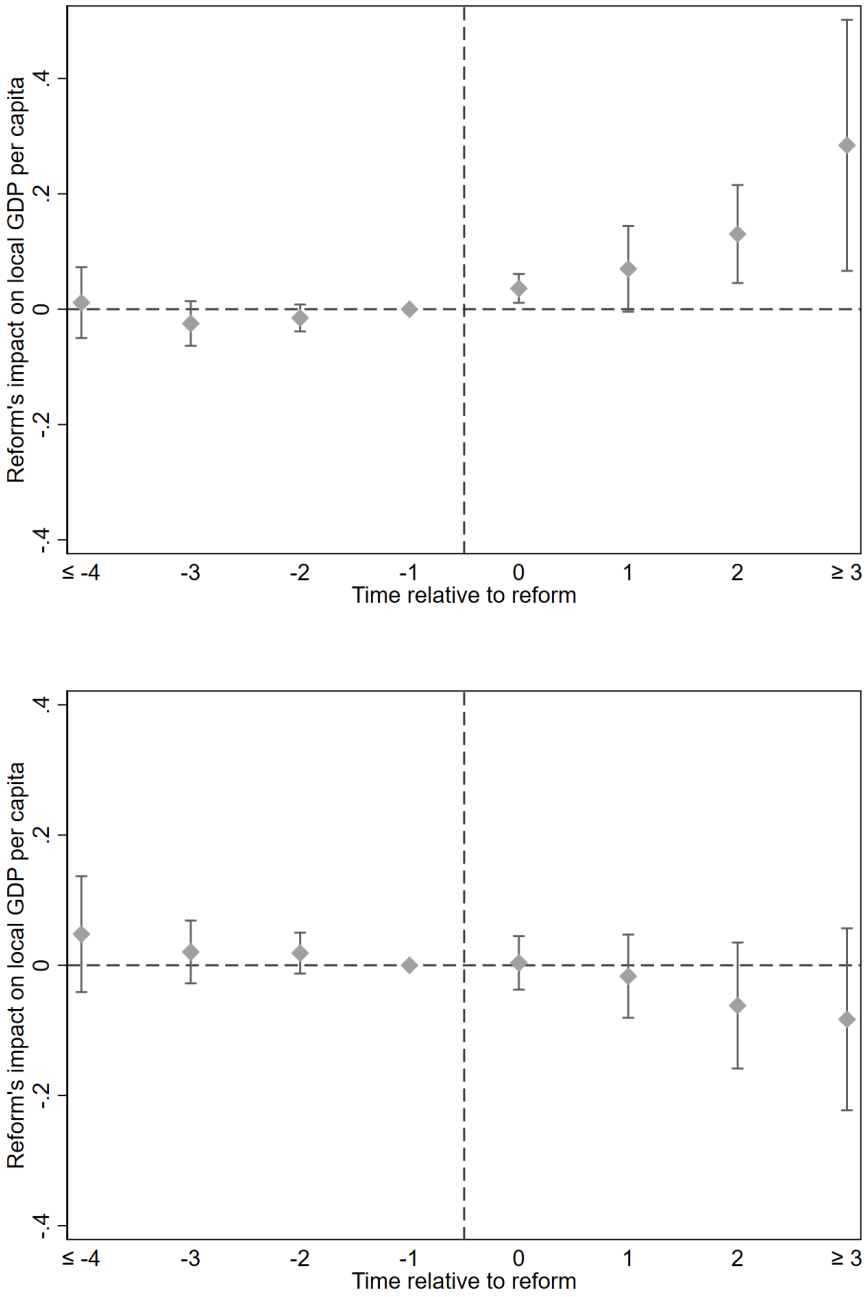


Figure A.17: Local fiscal reform - treatment effect on local GDP per capita. Upper Panel plots the counties reformed before 2007 and Lower Panel plots those reformed after 2007.

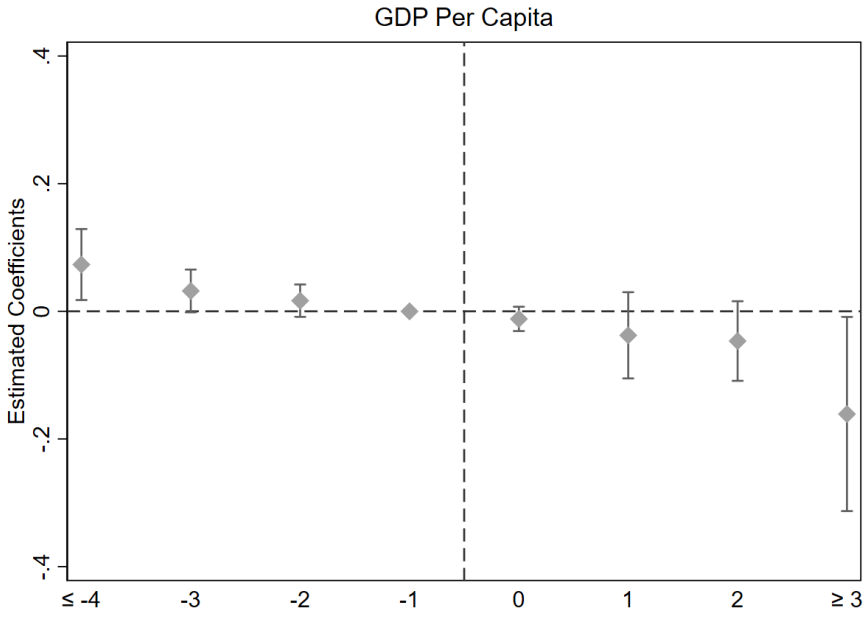


Figure A.18: Local fiscal reform - poor counties experimented before 2008. The pattern demonstrated here is similar to that in Figure A.17, Panel B.

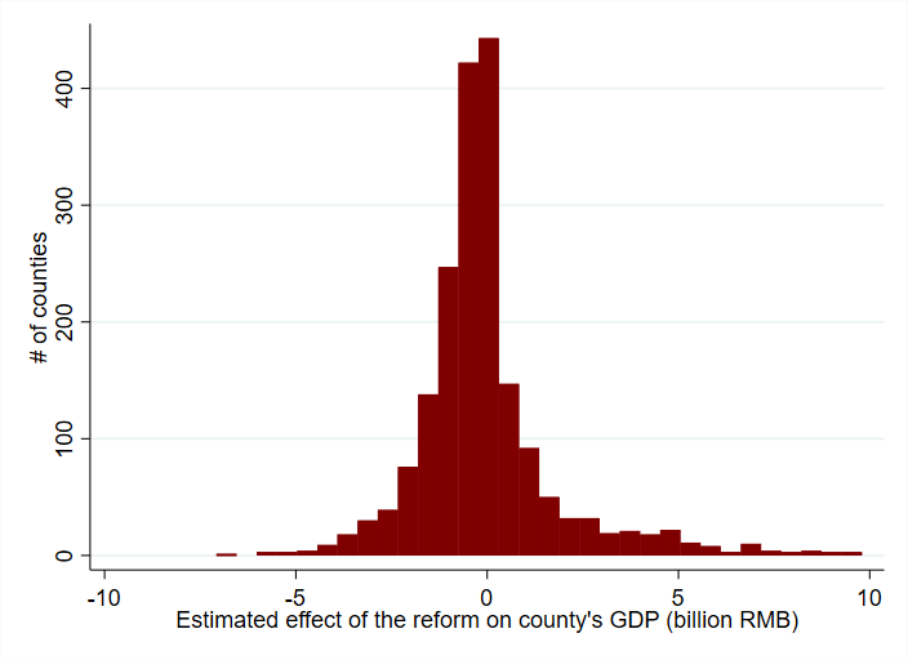


Figure A.19: Simulated treatment effects across country. We extrapolate the estimated treatment effect to all counties nationwide and obtain a distribution of reform effect on county's GDP.

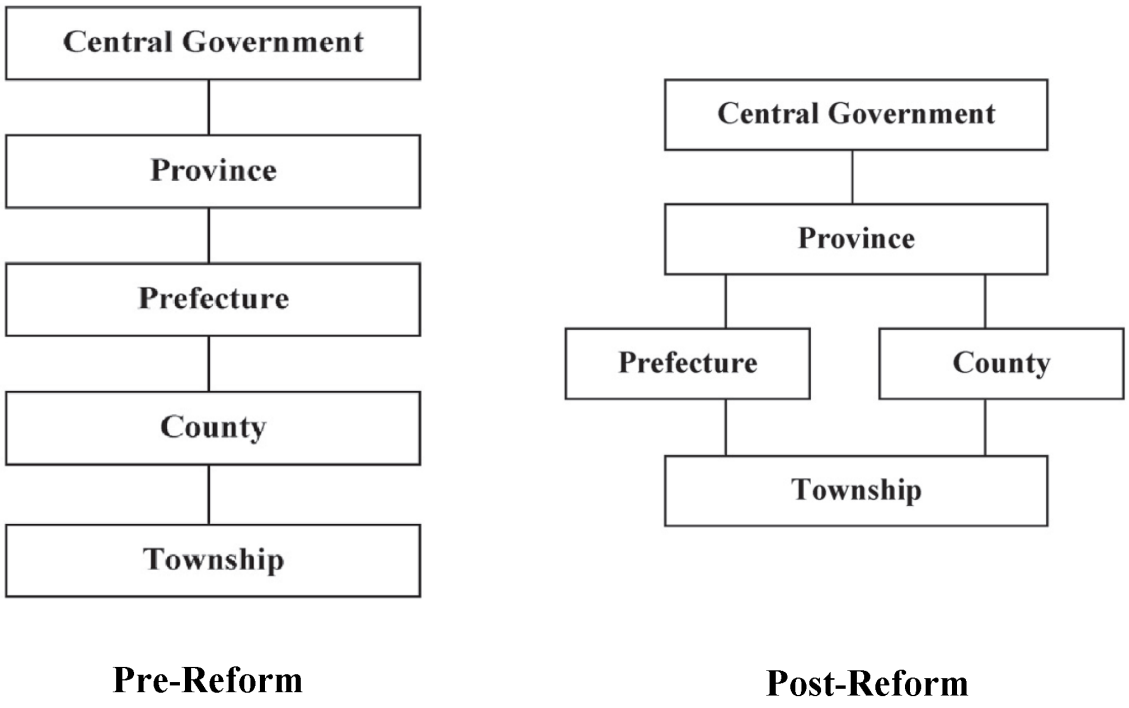


Figure A.20: Reproduced from Li et al. (2016). Illustration of local fiscal reform. After the reform, the provincial government could directly manage some of its counties, bypassing the prefectural cities, which grants county governments with more fiscal autonomy.

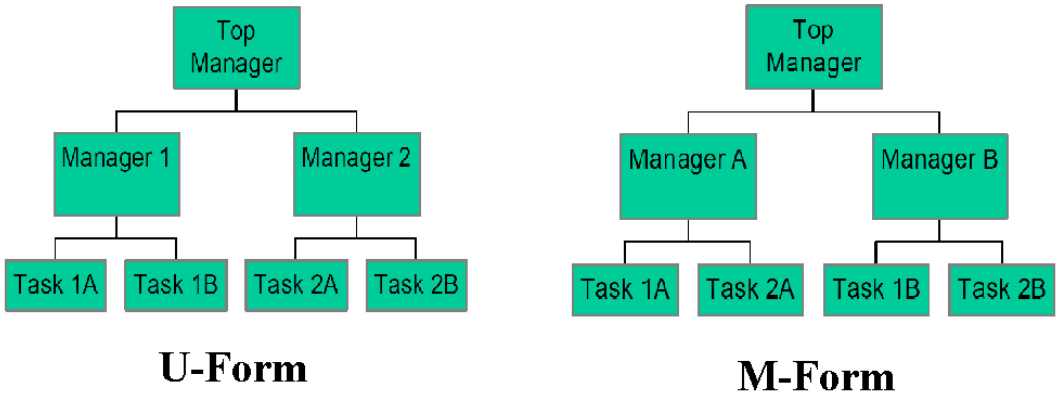


Figure A.21: Reproduced from Qian et al. (2006). Illustration of a shift from M form to U form. The top manager (ministers at central government branches in our case) have more administrative and personnel authority on its branches.

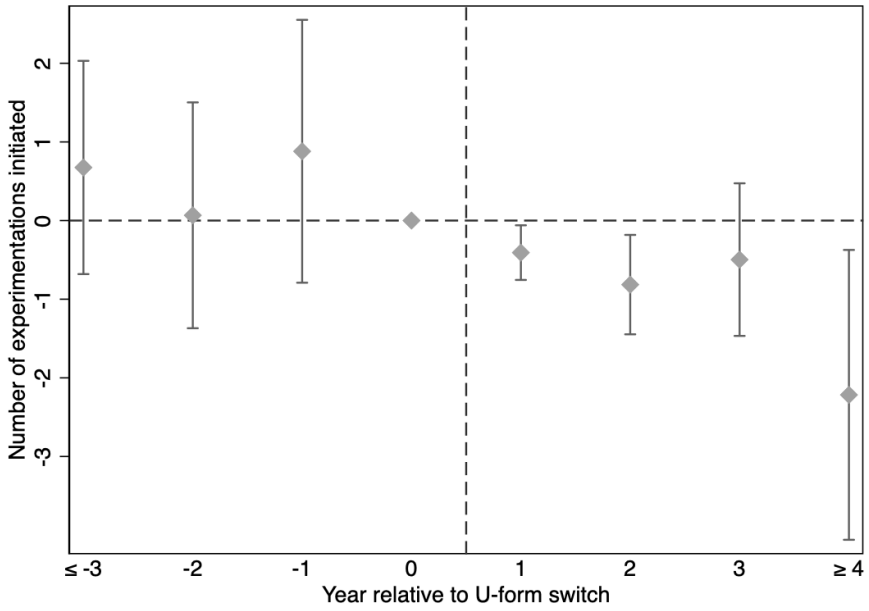


Figure A.22: Count of policy experimentations initiated after transitioning into U-form. X-axis indicates the time relative to the reform. The point estimates and confidence intervals are computed from a standard event study design controlling for ministry fixed effect and calendar year fixed effect. Standard errors are clustered at the ministry level.

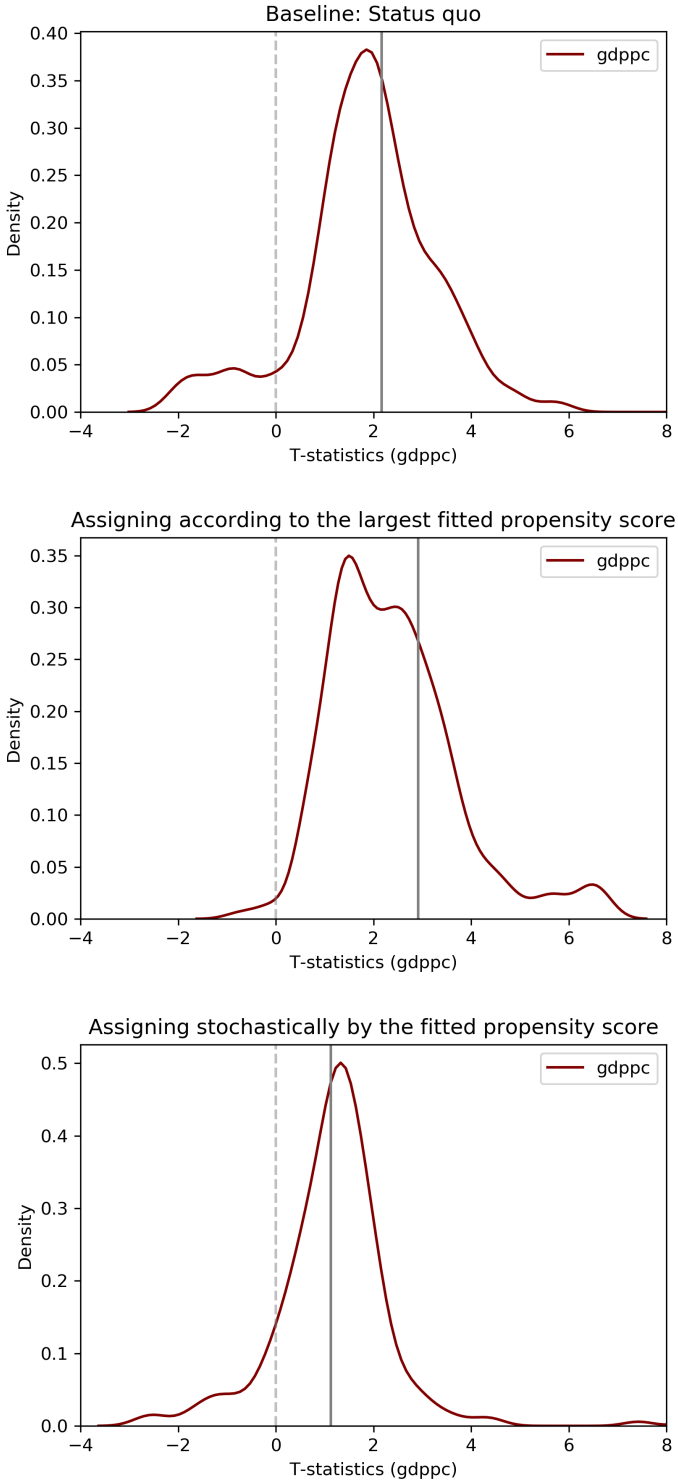


Figure A.23: Representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 2. To make sure we're making reasonable comparisons, we adjust the baseline Panel A so that it only includes the experimentations targeting prefectural cities, with 4 municipal cities excluded from observation. Panel B shows the simulated results when sites for experimentation are assigned to the prefectural units with the largest fitted propensity score. Panal C imposes a milder assumption where we assert a stochasitic version. Extreme values are trimmed. The grey vertical line indicates the mean of t-stats.

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

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

Note: In column (3) we report the ratio of experimentation-related policy documents issued by the central government that is found with its exact title in the *PKULaw* database. We then manually iterate through them. The numbers reported are very conservative. Fixing encodings of annotations and dropping secondary documents irrelevant to experimentation gives us a much larger ratio, but for consistency we do not report the calibrated numbers. In most cases, *PKULaw* collects even more documents than its official releases. For some ministries only recent years of documents are publicized on websites, (e.g. Fiscal and Tax; Natural Resources) so we confine the numbers we're comparing to the same time frame.

Table A.2: Political patronage and engagement in experimentations

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

Note: The standard errors clustered at the province level are reported below the estimates. We control for ministry by province fixed effect in all regressions. In column 2 we add time fixed effect and in column 3, we include the provinces' value added of first and second industry, fiscal expenditure and income of local governments as control variables. The mean of dependent variable and count of observations of Panel A alone are reported.

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

Table A.3: Political career incentives and engagement in experimentations

	Engaged in experimentations		
	(1)	(2)	(3)
Panel A: All experimentations			
Career incentive	0.397**	0.534**	0.469**
	(0.200)	(0.220)	(0.207)
Panel B.1: Experimentations initiated by M-form ministry			
Career incentive	0.407**	0.551**	0.492**
	(0.196)	(0.215)	(0.202)
Panel B.2: Experimentations initiated by U-form ministry			
Career incentive	-0.012	-0.017	-0.023
	(0.026)	(0.032)	(0.032)
Panel C.1: Experimentations with top-down assignments			
Career incentive	-0.023	0.059	0.030
	(0.106)	(0.127)	(0.119)
Panel C.2: Experimentations with voluntary sign-ups			
Career incentive	0.417***	0.475***	0.439***
	(0.136)	(0.142)	(0.138)
Mean of DV	1.146	1.146	1.146
Prefecture controls	NO	NO	YES
Politician controls	NO	YES	YES
Prefecture FE	YES	YES	YES
Time FE	YES	YES	YES
# of obs.	7117	7117	7117

Note: The standard errors clustered by prefectures are reported below the estimates. Control variables at the politician level include the educational level and previous central-government positions. Control variables at the prefecture level include GDP, GDP per capita, fiscal income, fiscal expenditure, and the ratio of added value by the third industry, all in logarithms. We also controlled for the career incentive of the previous city leader to address the concern where the engagement is just a continuation of previous progress. The construction of career incentive index is introduced in Appendix Section **??**= * panel A we report the estimated effect of career incentive intensity on all types of experimentation. In panel B we differentiate between experimentations issued by a M-form ministry, where city leaders have direct control on the logistics of the policy; and experimentations initiated by a U-form ministry where the central government takes direct orders on local branches. In Panel C, we investigate experimentations with top-down assignments and voluntary sign-ups, respectively.

Table A.4: Concerns for political stability and selection of experimentation sites

	Engaged in experimentations		
	(1)	(2)	(3)
Panel A: All experimentations			
# of protests in previous year	-0.003***	-0.002**	-0.002***
	(0.001)	(0.001)	(0.001)
Panel B.1: Experimentations initiated by M-form ministry			
# of protests in previous year	-0.003***	-0.002***	-0.003***
	(0.001)	(0.001)	(0.001)
Panel B.2: Experimentations initiated by U-form ministry			
# of protests in previous year	0.000*	0.000*	0.001***
	(0.000)	(0.000)	(0.000)
Panel C.1: Experimentations with voluntary sign-ups			
# of protests in previous year	-0.004***	-0.003**	-0.004*
	(0.001)	(0.001)	(0.002)
Panel C.2: Experimentations with top-down assignments			
# of protests in previous year	0.001	0.001	0.002
	(0.001)	(0.001)	(0.002)
Mean of DV	1.768	1.768	1.768
Pre-period controls	NO	NO	YES
Time FE	NO	YES	YES
Prefecture FE	YES	YES	YES
# of obs.	1547	1,547	773

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

Table A.5: Engagement in Experimentations and Local Government Expenditure

	(1) Log (Expenditure)	(2) Log (Expenditure)	(3) Log (Expenditure)
# of experimentations	0.060***	-0.008***	-0.002*
	(0.007)	(0.002)	(0.001)
Expenditure Category	Same	Irrelevant	Residual
Mean of DV	5.094	3.405	8.487
County-Year FE	YES	YES	YES
County-Category FE	YES	YES	YES
Year-Category FE	YES	YES	YES
# of obs.	159416	153458	153458

Note: The standard errors clustered at county level are reported below the estimates. This table shares the same specification with Table 4, but here we restrict our sample to policies not fiscally supported by the central government.

Table A.6: Engagement in Experimentations and Local Government Expenditure

	(1) Log (Expenditure)	(2) Log (Expenditure)	(3) Log (Expenditure)
# of experimentations	0.091***	-0.018***	-0.002
	(0.009)	(0.002)	(0.001)
Expenditure Category	Same	Irrelevant	Residual
Mean of DV	5.093	3.405	8.487
County-Year FE	YES	YES	YES
County-Category FE	YES	YES	YES
Year-Category FE	YES	YES	YES
# of obs.	159416	153458	153458

Note: The standard errors clustered at county level are reported below the estimates. This table shares the same specification with Table 4, but here we restrict our sample to policies not reliant on additional fiscal input.

Table A.7: Representativeness of experimentation sites selection and policy's national roll-out

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

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

Table A.8: Participation in local fiscal empowerment experimentation

	Engaged in experimentation		Emphasized locally	
	(1)	(2)	(3)	(4)
Ratio of "winner counties"	0.1146**	0.0863*	0.0026	0.0092**
	(0.0455)	(0.0519)	(0.0034)	(0.0039)
Mean of DV	0.565	0.565	3.039	3.039
Controls	No	Yes	No	Yes
# of obs.	1946	1937	1946	1937

Note: In columns 1 and 2, the dependent variable is a dummy indicating whether the county is eventually included in the experimentation; in columns 3 and 4, the dependent variable is the number of times that provincial government annual report emphasizes the roll-out of the reform. Ratio of "winner counties" represents the proportion of counties in a province that would benefit from the reform according to our simulation. Control variables include the number of villages in the county, land area and whether the county is a "winner" of the reform.

Table A.9: Similarity with experimentation sites and effects of policy roll-out: Robustness check

	(1) log_growth	(2) log_growth	(3) log_growth
Panel A. GDP per capita			
M-distance	-0.006***	-0.007***	-0.006***
	(0.001)	(0.001)	(0.001)
Panel B. GDP per capita + Fiscal income			
M-distance	-0.004***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)
Panel C. GDP per capita + Fiscal income + Population			
M-distance	-0.004***	-0.004***	-0.003***
	(0.000)	(0.001)	(0.001)
Observations	77,588	77,588	77,588
Policy FE	YES	NO	YES
County FE	NO	YES	YES
Mean of Dep. Variable	0.0806	0.0806	0.0806

Note: Robust standard errors clusters at policies are reported below the estimates. The idea of this table is illustrated in the notes of Table 8. The panel subtitles illustrates the different specifications of Mohalanobis distance we explored.

Table A.10: Non random experimentation and subsequent inequality

	Inter-county inequality			
	(1)	(2)	(3)	(4)
# policies originated from unrepresentative experimentations	0.003***	0.007**	0.004***	0.008***
	(0.001)	(0.003)	(0.001)	(0.003)
# policies originated from representative experimentations			-0.001	-0.001
			(0.001)	(0.001)
Mean of DV	0.395	0.395	0.395	0.395
Obs.	20	20	20	20
Important policies	No	Yes	No	Yes

Note: The 1998-2017 inter-county Gini index is given by $G_t = \frac{\sum_{i=1}^{m}\sum_{k=1}^{m}|y_{it}-y_{kt}|}{2m^2\bar{y}_t}$. Important policies are defined by those mentioned in the Five-year plans.

Table A.11: Ministries underwent vertical management reforms

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

Table A.12: Predicting politicians' career incentives

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

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

Table A.13: Political incentives and engagement in experimentations

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

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