The Political Economy of Policy Experimentations in China

Shaoda Wang
David Y. Yang*

April 2, 2021

Abstract

Policy learning is an essential component of government's decision making. Many have speculated that systematic policy experimentations play a vital role in fostering China's reform and growth. In this project, we aim to describe and understand China's policy experimentations. We collect comprehensive data on policy experimentations conducted in China over the past 4 decades by 139 ministries and commissions. We document four main results. First, flat, flexible, and decentralized state organizational structure facilitates policy experimentation; as the ministries undergo reform from M- to U-form, the number of policy experimentations they administer declines. Second, among the policy experimentations that do take place, a substantial share deviate from representative experimentation sites selection as they over-sample more developed jurisdictions. Third, deviation from representativeness carries significant information cost, and biases policies to systematically favor more developed regions. Fourth, about half of the deviation from representativeness can be attributed to misaligned incentives between the central and local governments. 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 Tim Besley, Mike Callen, and Jesse Shapiro for stimulating comments and suggestions. Weicheng Cai, Andrew Kao, Kaicheng Luo, Jiarui Qian, and Bobing Qiu provided outstanding research assistance.

1 Introduction

Determining which policies to implement is an essential component of government's decision making (e.g., Hayek, 1978; North et al., 1990; Roland, 2000; Mukand and Rodrik, 2005). Sophisticated policy learning could involve sequences of trials and errors. Several notable government initiatives even include components of rigorous randomized control trails, influenced by one of the most notable methodological innovations in the discipline of economics over the past two decades (Banerjee and Duflo, 2009).

While many governments have explicitly or implicitly engaged in policy experimentations in various forms, few 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 policy reforms across regions and often in multiple waves, and such experimentations help the government discover and implement the policies that yield desirable outcomes. In fact, many scholars have speculated that policy experimentations were vital to China's growth over the past four decades (e.g., Rawski, 1995; Cao et al., 1999; Roland, 2000; Qian, 2002).

This project aims to describe and understand China's policy experimentations since the 1980s. Specifically, we aim to answer four questions. First, under what circumstances do policy experimentations take place? Second, among the experimentations that do take place, are the experimentation sites chosen in a manner that is representative of the country? Third, if the site selection is not representative, what are the consequences of such deviation from representativeness in policy experimentations? And finally, what factors drive experimentations' deviation from representativeness?

To answer these questions, we collect comprehensive data on policy experimentations in China between 1980 and 2020. Based on 19,644 government documents, we construct a database of 854 policy experimentations initiated by 139 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 construct a variety of characteristics of policy experimentations based on the associated government documents, as well as other linked datasets such as the bureaucratic structure of the policy-initiating ministries, local socioeconomic conditions, and career trajectories of central and local politicians.

We begin with investigating the overall trend of policy experimentations, and the conditions under which policy experimentations take place. We find that there has been a steady increase in the total number of policy experimentations implemented since 1980

when the Reform-and-Opening-Up era began, and the speed of introduction of new experimentations picked up in the 1990s. Since the early 2010s, however, the number of experimentations sharply declined. 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, funding, 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). Using an event study design, we demonstrate that following the transition to U-form organization, the vertically managed ministries significantly decrease the amount of policy experimentations they administer. 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.

We next focus our attention on the policy experimentations that do take place, and examine how the experimentation sites are selected. 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, when we compare the pre-experimentation characteristics between the locations that are selected as experimentation sites and those that are not, we observe that majority of the experimentations are conducted in sites that are positively selected in terms of local economic conditions. For at least 40% of the policy experimentations, we can reject the null hypothesis of representative site selection at the 90% confidence level.

We then turn to the consequences of deviation from representativeness in policy experimentations. Deviation from representativeness would imply information loss if there exists substantial heterogeneity in the mapping from policy to outcome across localities. We illustrate such information loss, first zooming in a specific policy experimentation on local fiscal empowerment. In order to foster the economic growth across Chinese counties, the central government initiated an experimentation between 2003 and 2013 that provides fiscal empowerment and even fiscal autonomy to over 1,100 counties, allowing provincial governments to bypass prefectural governments and directly administer the experimented counties within their jurisdictions. We observe that the sites are substantially positively selected at the beginning (t-stat of a t-test on pre-experimentation GDP per capita between experimented and non-experimented counties is as high as 9.110 during the first two years of the experimentation), then become negative selected (t-stat

decreases to -6.156 by 2007), and have shifted towards representative selection by the end of the experimentation. Using a staggered event study design controlling for location and time fixed effects, we find considerable heterogeneity in the effects of this experimentation on local economic development: counties with higher pre-experimentation GDP per capita benefited from the experimentation, while the poorer counties experienced worsened subsequent local economic development. The non-representativeness of the experimentation, in particular the positive selection at the early stage of the experimentation that disproportionately highlights the positive effects of the experimentation, masks the unequal nature of the fiscal empowerment policy and could bias the central government's policy choices. In fact, had the experimentation been rolled out to the entire country, it would generate a net zero effect with both winners and losers.

Zooming out of the fiscal empowerment experimentation, we next examine the overall consequences of unrepresentative experimentations. If the unrepresentative (in particular, positively selected) experimentations tend to cast a positive light on the policies that disproportionately benefit richer localities, as illustrated in the example of the local fiscal empowerment policy, then such deviation from representative experimentation could systematically bias the space of implemented policies and reforms in China, and generate distributional consequences across regions. We observe that experimentations that are closer to representativeness in their implementation 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. Moreover, 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 benefit substantially more from the policies. Indeed, linking the number of experimentations that have successfully rolled out to the entire nation in a give year to the subsequent inter-county inequality in economic development, we find that a higher number of national policies originated from unrepresentative experimentations is associated with significantly *higher* subsequent inter-county inequality. Importantly, this is not the case for national policies that originated from experimentations closer to representativeness — if anything, those policies are associated with a lower subsequent level of inter-county inequality.

Finally, we investigate what cause the policy experimentations to deviate from representativeness. 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. We then document patterns that suggest potential misaligned incentives between the central and local governments in generating nationally-representative knowledge about policy effectiveness: experimentations are closer to representativeness if the site selections are assigned by the central government directly rather than involving voluntary participation by the local government, and the deviation from representativeness is not nearly as severe at the provincial level, as compared to the choices of specific prefectures and counties to be the experimentation site. We further explore the precise political constraints between the central and local governments 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 rooms 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. Furthermore, 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. Overall, the misaligned incentives between central and local governments as well as the central government's demand for political stability could account for more than half of the positive selection in experimentation sites we observe.

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 institutional (both political and organizational) conditions and contexts. Such policy experimentations will be implemented with a set of the institutional and bureaucratic conditions that may serve as the initial engine that motivates and organizes policy experimentations. However, as we show, such institutional and bureaucratic contexts also imply the presence of constraints, such as incentive misalignment within the polity. These could be forces that push against experimentation both on the extensive margin (e.g., how many experimentations can happen in the first

place), and on the intensive margin (e.g, how well is centrally planned experimentation design such as representativeness adhered, and how much local differentiation there is among the experimentation participants).

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 our context, 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.[1]

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 between 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 en-

[1]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.

vironment stand in contrast with theoretical works analyzing experimentations in decentralized, voluntary contexts (Mukand and Rodrik, 2005; Callander and Harstad, 2015).[2] 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.[3]

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 could gather during the policy learning process.[4]

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

[2]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.

[3]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).

[4]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).

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).[5]

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 and the process to construct the experimentation database. Sections 4 to 7 answer each of the four questions that we raise, and Section 8 concludes.

2 Institutional background

China's economic decentralization and reforms are accompanied by its high-powered political centralization (Xu, 2011). As a result, the central government plays a key role in initiating and coordinating policy experimentations. Such a top-down approach of policy experimentation is described by Heilmann (2008b) as "experimentation under hierarchy." This stands in contrast with the spontaneous policy experimentations that often take place in federalist polities (Shipan and Volden, 2006; Cai et al., 2009; Callander and Harstad, 2015).

China's tradition of policy experimentations started with Mao Zedong in the 1940s during the Communist Revolution (e.g., through implementation of land reform in selected regions in order to consolidate the Communist regime). 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). 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).

[5]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).

In this section, we briefly describe the forms and administrative procedures in which policy experimentations typically take place. We also introduce two particular policy reforms that we examine in depth in the paper — the government vertical management reform and the local fiscal empowerment reform.

2.1 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). This form of policy experimentation is the primary focus of the empirical analyses of our paper.

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 (i.e., 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).[6] Policy experimentations implemented in the experimental zones are included in our analyses, though they are often categorized as voluntary (bottom-up) experimentations rather than centrally assigned (see Section 3.1 for details).

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, 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.[7]

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. Such form of

[6]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 reforms during the Reform-and-Opening era.

[7]Source: Zhou (2013), page 130.

policy experimentation is the most prevalent as well as the most well-defined: experimentation points has clear regional variation in experimentation status; and for each policy, the corresponding ministry is clearly set out to use a certain number of experimentation sites to shed light on the costs and benefits of the policy before its national roll-out.

2.2 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.

2.3 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 Section 4.2.

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.1 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.1 for a list of the ministries that underwent the vertical management reforms and the years at which they took place.

2.4 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 in Section 6.1.

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).[8] 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.2, 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.[9]

Within a decade, more than 1,100 counties in China were assigned as the experimentation sites of the reform. The experimentation was rolled out in multiple waves. Based on the central government's document that guides the fiscal empowerment reform, we

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

[9] 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).

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. As we will demonstrate in Section 6.1, 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.

3 Data

We compile, to the best of our knowledge, the most comprehensive dataset on policy experimentations in China over the past four decades. We match the policy experimentations to several other datasets, such as the ones that describe government organizational structure, backgrounds of involved politicians, and regional socioeconomic conditions.

3.1 Government documents on policy experimentations

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

From *PKULaw.com*, we collect the (nearly) universe of government documents between 1980 and 2020 containing the key word "experimentation points" (*Shidian*). We obtain 19,644 documents in total, among which 4,231 are issued by the central government, and 15,413 by the local government branches. Based on policy themes, these 19,644 government documents can be categorized into 854 distinct policy experimentations, these policy experimentation involve a total of 139 different central government ministries and commissions.

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 we find that the government documents collected from the *PKULaw.com* has extensive and comprehensive coverage (see Appendix Table A.2). Importantly, we manually examine the limited documents that are published on the ministries' websites but not included in the *PKULaw.com* database, and we find that they are secondary documents and do not

contain information on policy experimentations.

By relying government documents to describe policy experimentations, the experimentations need to reach a stage of formal endorsement and initiation by the central government in order to be included in our sample. Thus, we do not observe very early stage experimentations started by the local government that never reach to the level of the central government. While we cannot direct evaluate the magnitudes of such early stage experimentations, we observe that 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.

We extract several key pieces of information from the corresponding government documents in order to characterize each policy experimentation. First, 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). Second, since many policy experimentations have more than one wave of roll-out, we extract the roll-out schedule of each policy experiment, based on which we identify 1,715 distinct rounds of roll-out for the 854 policy experimentations. We link each scraped government document to a specific policy-round based on its date of issuance, which allows us to compare site-selection patterns across rounds. Third, we collect detailed information on the localities participated in policy experimentations, for example, the timing of their joining the policy experimentation. Fourth, based on whether the roll-out eventually covered at least two-thirds of the whole country, we can define whether a policy experimentation as either successful or failed.[10] 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. Sixth, we measure the importance of specific policy experimentations by whether they are explicitly mentioned in the central government's Five Year Plans documents.[11] We proxy for the complexity of the 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. Finally, we measure how policy innovation and differentiation evolved across time and space, by construct matrices

[10]The definition of successful and failed policy experimentations is not sensitive to the exact roll-out cutoff that we use. We also cross-validate this definition by checking whether the experimented policy eventually got formalized in subsequent government documents.

[11]Some policy experimentation documents also mentioned the stated goal of the experimentation, and whether the central government would provide additional fiscal support for the experimentation sites. Such information was collected when available.

of pairwise textual similarities for all the local policy documents that belong to the same policy experimentations.[12]

3.2 Auxiliary data sources

We match our dataset on policy experimentations with three additional sources of data. First, 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.[13] 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.[14]

Second, 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. We discuss the construction of this proxy in greater detail in Appendix B.

Third, 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.

Finally, 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

[12]Text similarity is calculated using *Paddluehub*, an open-source natural language processing model pretrained by *Baidu*. We use a convolutional neural network for the representation layer, and MatchPyramid model for the semantic layer, the network configuration is set up point-wise. See Paddle's Github repository for details.

[13]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.

[14]For ministries that merged with each other, the unit of analysis is the eventually merged ministry throughout the sample period.

databases on global political events.[15]

3.3 Summary statistics

We present basic summary statistics of the policy experimentations in Table 1: Panel A lists the total number of policy experimentations initiated during 1980 and 2020, the average number of rounds and experimentation sites involved in each experimentation, as well as the overall success rate of the experimentations; and Panel B further divides the experimentations based on policy domains.

4 Under what circumstances do policy experimentations take place?

4.1 Trend of policy experimentations over time

We begin with the first question that we ask: under what circumstances do policy experimentation take place? We first examine the trend of the amount of policy experimentations initiated in China since 1980.

Figure 1 plots the total number of policy experimentations implemented over time. We record the first year when a specific policy experimentation started as the year associated with the multi-year roll-out of the experimentation. One observes that 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.

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. If we count each round of policy experimentation as a separate experimentation, and examine the number of experimentation rounds introduced in each year (see Appendix Figure A.3 for a plot of the count over time), we observe a similar pattern that the amount of experimentations increased sharply until early 2010s, and then dramatically declined afterwards.

[15] See `www.gdeltproject.org` for details of the GDELT Project.

We observe similar hump-shaped pattern when we separately examine the experimentations that primarily rely on the central government assigning experimentations sites and those that involve (at least to some degree) voluntary participation of local experimentation sites. Appendix Figure A.4 separately plots the counts of these two types of experimentations over time.

Despite the robust pattern of the overall time trend in policy experimentations initiated by the central government, there exists substantial heterogeneity across the government ministries and commissions. We group ministries and commissions by the broad functions that they are responsible for, and Appendix Figure A.5 plots the count of policy experimentations belonging to each ministerial functions over time. The Ministry of Education and the Ministry of Agriculture initiated the most amount of policy experimentations in terms of total volume, ranging from 2.3 to 2.1 per year, respectively.[16] The policy experimentations from these policy domains also have different time trends: while experimentations regarding commerce and business experience a similar decline after early 2010s as the rest of the experimentations, experimentations issued by the State Council and Reform Commission remain steadily increasing throughout the period.

4.2 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

[16] In cases where a particular policy experimentation is introduced by multiple ministries and commissions, we identify the primary ministry or commission that leads the experimentation.

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 2.3, 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:

$$y_{mt} = \sum_k D_{mt}^k \cdot \beta_k + \delta_m + \theta_t + \varepsilon_{mt}, \tag{1}$$

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

through the M- to U-form transition in different years. The baseline specification clusters the standard errors at the ministry/commission level.

Figure 2 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.[17] 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.

5 Are experimentation sites representative?

We next focus our attention on the policy experimentations that do take place, and examine the selection of experimentation sites. From the central government's perspective, a key criterion for experimentation site selection is its representativeness (see Section 2.1 for details). As a benchmark, we thus examine whether the selection of experimentation sites is indeed representative.

[17] 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.

5.1 Procedure to test for representativeness

Specifically, for each policy experimentation, we compare pre-experimentation characteristics between locations where experimentation is implemented ("treatment" sites) and those that do not undergo the experimentation ("control" sites). As the baseline, we examine the local GDP per capita during the year prior to experimentation roll-out, and we discuss below the robustness of using a variety of other local characteristics.

For each experimentation, we conduct a t-test against the null hypothesis that the pre-experimentation local GDP per capita is indistinguishable between the experimentation sites and non-experimentation sites. This amounts to 854 independent t-tests, one for each policy experimentation.

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

$$t = \frac{\hat{Y}(1) - \hat{Y}(0)}{\sqrt{\frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}}}, \tag{2}$$

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

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}. \tag{3}$$

We also derive indicators, for each policy experimentation, whether we can reject the null hypothesis that the experimentation sites are representative of the country at the 10% confidence level. Note that drawing a single confidence band here does *not* mean we are conducting a uniform balance test across all polity experimentations. For each policy experimentation and its 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.

The specific context of China's policy experimentations poses two important complications to conducting the representativeness tests. First, policy experimentations are implemented at different administrative levels, and thus the experimentation site selections are made at distinct levels. Vast majority of the experimentations fall under two categories: those that are conducted at the provincial level, and those conducted at the county level. The latter represents 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. We con-

duct 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 within the same province.

Second, approximately one-fourth of the policy experimentations only involve one experimentation site, although the average number of experimentation sites are 17.36 across all experimentations. There's no way to derive a variance estimate from one sample point, so we cannot conduct the 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 control locations are defined as those that have never been a site for experimentation across the five policies. 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 for 10,000 times. We discuss the robustness of our baseline findings with respect to these alternative specifications below.

The tests described above, in particular, conducting representativeness tests for each policy experimentation separately, is a conservative one. If we were to identify deviations from representativeness with these tests, then a pooled test with multiple experimentations at once would yield more power in detecting unrepresentativeness and reject the null hypothesis.

5.2 Test results

We plot the distribution of the t-statistics, pooling all policy experimentations together, in Figure 3, Panel A; Panel B presents the corresponding cumulative distribution function plot. In both figures, we mark the thresholds of t-statistics where one would reject the null hypothesis of representative experimentation site selection.[18] Table 2, Panel A, reports the corresponding test statistics and the share of policy experimentations that one can reject the null hypotheses at the 5% level.

[18] As discussed above, there should be around 850 mutually independent t-tests that all have their respective degrees of freedom. Some are wider, and some narrower. We choose 2.57 because the average localities involved in any given experimentation is somewhere around 6, and the 97.5 quantile of t distribution with degrees of freedom 5 is 2.57 (when $n_1 \ll n_2, \nu \to n_1 - 1$). Precise tests adjusting for the degree of freedoms, or generated by permutations, are reported in Table 2 as well as the subsequent tables.

When we compare the pre-experimental characteristics between the experimentation sites and locations not selected for experimentation for each policy experimentation, we are able to reject the null hypothesis of representative selection among at least 40% of the experimentations at the 10% level. This is a conservative test: conducting such tests at the level of each individual policy experimentation yields statistically conservative null hypothesis rejection rates. In fact, the majority of the experimentations are positively selected in terms of pre-experimentation local GDP per capita.

The magnitude of the deviation from representative selection and the underlying positive selection in sites is remarkably robust. First, we observe similar patterns 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.6, Panels A to C, respectively). Second, the patterns are similar when we separately examine the multi-site experimentations — among which tests are conducted independently for each experimentation — and one-site experimentations — among which tests are conducted by pooling consecutive experimentations together (see Appendix Figure A.7, Panels A and B, respectively). Third, it is also robust to alternative permutation tests among the multi-site experimentations (see Appendix Figure A.8), 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.9). Finally, the pattern that we observe is robust if we just examine the subsample of early-round experimentation sites, thus effectively holding fixed the number of experimentation sites across policy experimentations and do not allow the extent of subsequent experimentation roll-out to affect the representativeness test of initial implementation (see Appendix Figure A.10).

While we observe a relatively stable share of experimentations deviating from representativeness over time, there exists substantial heterogeneity across policy domains: ranging from tax and fiscal policies and sciences and education policies on one end of the spectrum, both are substantially positively selected, to agricultural policies on the other end of the spectrum, whose experimentation sites are only mildly positively selected (see Table 2, Panel B).

6 What are the consequences of unrepresentative experimentations?

We next turn to the consequences of such deviation from representativeness in policy experimentations. 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.[19]

We illustrate such information loss first using the context of a specific experimentation on local fiscal empowerment in Section 6.1, then we zoom out and examine the overall consequences of unrepresentative experimentations in Section 6.2.

6.1 Information loss due to unrepresentativeness: a case study of the local fiscal empowerment experimentation

The local fiscal empowerment 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.

As we introduced in Section 2.4, 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 first examine whether the site selection of this particular experimentation is representative, following the strategy introduced in Section 5.1. We treat each year of the decade that the local fiscal empowerment experimentation rolls out as a separate experimentation, comparing the previous local GDP per capita in counties that are included as an experimentation sites with those that have not yet participated in the experimentation up until the corresponding year.[20] We plot the t-statistics for representativeness for each year in Figure 4, where we stratify the test at the province level to take into account the fact that the choices of experimentation counties are made by the provincial governments who participate in the experimentation (results are similar if we do not stratify by province, as shown in Appendix Figure A.11).

We observe that the experimentation sites are substantially positively selected at the

[19]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, as suggested by Rogger and Somani (2018); Hjort et al. (2019); Vivalt and Coville (2019).

[20]Specifically, Within each province, we consider counties that engage in the experimentation for the first time as the units of the treatment group, the rest as control. Once a county had participated in the reform, it would not be included in the control group in all subsequent years.

beginning of the experimentation (t-stats of a t-test on pre-experimentation GDP per capita between experimented and non-experimented 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.

Next, we demonstrate the existence of a necessary condition for information loss due to unrepresentative experimentation: heterogeneous treatment effects along the dimensions that the experimentation deviates from representativeness. Specifically, we use the following 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):

$$y_{pct} = \sum_k D^k_{pct} \cdot \beta_k + \delta_c + \gamma_{pt} + \varepsilon_{pct}, \tag{4}$$

where y_{pct} is the local GDP per capita in county c of province p at year t, and D_{pct} is the indicator of years relative to county c's implementation of the local fiscal empowerment experimentation. We control for a full set of county fixed effects (δ_c), and province-specific time fixed effects (γ_{pt}) to account for regional heterogeneous time trends in economic development. The results are presented in Figure 5, where Panel A shows the treatment estimation for early rounds experimentation counties, and Panel B for experimentation counties during the later rounds.

We find considerable heterogeneity in the effects of such experimentation on local economic development: counties that have higher pre-experimentation GDP per capita benefited from the experimentation, while the poorer counties experience worsened subsequent local economic development. Importantly, such patterns of heterogeneity by pre-experimentation local economic conditions is *not* merely reflecting 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.12).

The experimentation, had it been rolled out to the entire country, would generate a net zero effect with both winners and losers. We plot, in Appendix Figure A.13, the distribution of the projected treatment effects of the local fiscal empowerment for each county in China. One observes that while there is a long right tail, suggesting that a handful of counties substantially benefit from local fiscal empowerment, the distribution centers around zero and there is a large mass of counties with negative treatment effects from the

exact same policy experimentation.

Moreover, 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.3).

The 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.

6.2 Unrepresentative experimentations, information loss, and inequality-inducing policies

We now investigate the overall consequences of unrepresentative experimentations, examining whether the local fiscal empowerment experimentation represents a more general phenomenon arising from the overt positive selection of policy experimentations.

To the extent that positively selected policy experimentations could generate a structural bias towards representing and deriving national policies that might disproportionately benefit richer localities, such deviation from representative experimentation could systematically bias the space of implemented national policies and reforms in China, and generate distributional consequences across regions.

We first 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.4 presents results where we predict the likelihood of policy roll-out based on the corresponding policy experimentation's being unable to reject from the null hypothesis of a representative experimentation (columns 1-4) or the associated t-statistics from the representativeness tests (columns 5-8). We observe a robust pattern that deviations from representativeness are associated with higher chances of rolling out to the entire country.

This reflects the central government's desire to learn about the mapping from policy to outcomes, and the underlying uncertainty in such mapping — representative experimentations are more likely to reveal sub-optimal policies. On the contrary, unrepresentative, in particular, positively selected policy experimentations may appear to be (artificially) successful and effective in the early rounds of experimentations, as illustrated in the example of local fiscal empowerment policy.

Moreover, when experimented policies roll out to the entire country, localities that share similar socioeconomic conditions with the sites where the corresponding policies were experimented on benefit substantially and differentially more from the policies. Specifically, for each policy and locality that did not participate in the experimentation, we calculate a Mohalanobis distance, M_{cp}, between its vector of pre-experimentation GDP, local fiscal income and fiscal expenditure, and the matrix of corresponding characteristics of all localities in the experimentation. We then estimate the relationship between local GDP growth after policy roll-out among localities that do not participate in the experimentation and their distance to experimentation sites, controlling for locality fixed effects (γ_c) and policy fixed effects (η_p):

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

Table 3 presents the results. One observes that localities more different from the experimentation sites benefit *less* from the policies when they roll out to the rest of the country beyond the experimentation sites. This is predominantly driven by policies in the economic and development domains (see column 3).

These results suggest that policies originated from unrepresentative experimentations differentially benefit some regions over others. We next investigate whether there indeed exists a relationship between policies originated from unrepresentative experimentations and the subsequent inter-regional inequalities in China. We calculate the inter-county Gini coefficient in GDP per capita by:

$$G_t = \frac{\sum_{i=1}^{m} \sum_{k=1}^{m} |y_{it} - y_{kt}|}{2m^2 \bar{y}_t}, \tag{6}$$

and we focus on the policy experimentations that have rolled out to the entire country and become national policies.

Table 4, column 1, presents the results where we link the inter-county Gini coefficient in economic development with the number of policies implemented in the previous year that are derived from unrepresentative experimentations. One observes that a higher number of national policies originated from unrepresentative experimentations is asso-

ciated with significantly *higher* subsequent inter-county inequality in economic development. Such association between unrepresentative-experimentation-induced national policies 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* (see column 2).

On the contrary, this is not the case for national policies that originated from experimentations that are closer to representativeness in their implementation (see columns 3 and 4). If anything, those policies are associated with a lower subsequent level of inter-county inequality. Therefore, it indeed appears to be the case that positively selected policy experimentations produce a portfolio of policies that are in favor of certain regions at the expense of others, thus leading to greater inter-regional inequality throughout China.

7 What drives deviation from representativeness?

Given the consequences of unrepresentative experimentation, we next investigate the factors that may drive the observed deviations from representativeness. In Section 7.1, we first examine whether the observed positive selection can be justified by optimal experimentation design with alternative motives of the decision makers such as maximizing their subjective expected utility and welfare of the experimentation sites. In Section 7.2, we investigate factors related to misaligned incentives between the central and local governments that could induce deviation from representativeness when experimentations are implemented. Finally, in Section 7.3, we discuss other objectives held by the central government when it selects experimentation sites that may contradict representativeness.

7.1 Can we justify the observed positive selection?

If 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, then unrepresentative roll-out of experimentations may be justified. We now 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.

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 esti-

mated heterogeneous treatment effects from Section 6.1; Appendix A.1 provides details on the simulation procedure. 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.

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 6.1; Appendix A.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 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.

7.2 Misalignment between the central and local governments

If one cannot justify the severe positive selection as optimal experimentation design, could such positive selection occur even if the central government genuinely intends to conduct representative experimentations?

Two aspects of heterogeneity in the representativeness of policy experimentations suggest that it could be the case. 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 Appendix Figure A.14). 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 B). These 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. Indeed, it has also 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), or to help minimize the negative impacts of a failed policy experiment on a poor region (Wu, 1995; Huang, 2000). We discuss the political sources of such misalignment more explicitly in the next section.

Since centrally assigned experimentation sites appear to be more representative than locally-assigned or locally-volunteered sites, it is likely that there exists potential misalignment between the central and local governments' incentives, which drives the deviation from representative policy experimentation. In this section, we directly investigate the political incentives behind policy experimentations.

7.2.1 Political patronage in policy experimentation

Existing literature has documented the salient role of political patronage in China's government system Fisman and Wang (2015); Fisman et al. (2020). Given the potential political rewards associated with successful policy experimentations, it would not be surprising that political patronage also serves as a contributing factor to the unrepresentative selection of experimental sites.

To investigate this hypothesis, we exploit the inter-temporal changes in a region's connection to each ministry created 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 can't influence the appointment of central ministers, the turnover of ministers can be regarded as exogenous shocks to the province-ministry connections. Specifically, we estimate the following econometric model using ministry-province-year level data:

$$y_{mpt} = \alpha \cdot connection_{mpt} + X'_{pt} \cdot \beta + \delta_{mp} + \theta_t + \varepsilon_{mpt}, \tag{7}$$

where y_{mpt} is whether ministry m assigned a policy experiment to province p 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.

δ_{mp} represents province-by-ministry fixed effects, θ_t represents year fixed effects. The standard errors are clustered at the policy level.

As shown in Table 5, Panel A, when a region gets connected to a minister, it is suddenly 3.7% more likely to be assigned as an experimentation site by that ministry. The effects are completely 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.5). This indicates that the political patronage in experimentation site assignment likely works through top-down favoritism.

To visualize the dynamic impacts of political patronage, in Appendix Figure A.15, we present the event study estimates exploiting minister turnovers: in years prior to being connected to a ministry, a region's likelihood of being assigned as experimentation site by that ministry remained stable over time; but once the region gets connected to the ministry, it is immediately more likely to be assigned as an experimentation site by that ministry. The flat pre-trend is reassuring, as it suggests that "being connected to a ministry due to central minister turnovers" is indeed likely orthogonal to the counterfactual trajectories of local governments' experimentation behaviors.

7.2.2 Political career incentives and engagement in policy experimentation

We next examine how the prefectural city leader's career advancement incentive affects his participation in policy experimentations, by exploiting within-prefecture changes in leader's political incentives.

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 his age, tenure, and official rank in the bureaucratic system (see Appendix B for details of the construction of this measure). This measure of upward mobility can be used as a proxy for the politician's career advancement incentives: when the politician is more likely to be promoted in a certain year, he will have stronger incentives to demonstrate successes and achievements to help build his promotion case.

Specifically, we estimate the following econometric model:

$$y_{pt} = \alpha \cdot incentive_{pt} + X'_{pt} \cdot \beta + \delta_p + \theta_t + \varepsilon_{pt}, \tag{8}$$

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; X'_{pt} is a vector of time-variant regional control variables; δ_p and θ_t represent region fixed effects and year fixed effects, respectively. Standard errors are clustered at the prefectural

city level.

As shown in Table 5, Panel B, when the local prefectural city leader has strong promotion incentives, the prefectural city indeed engages in significantly more policy experimentations. Given the inclusion of prefectural city fixed effects and year fixed effects, the identifying variation comes from within-prefecture changes in leader's political incentives, which is generated either by leader transition over time, or by the same leader getting more/less incentivized as he ages. In either case, the identifying variation is likely exogenous to the initial trend of each prefectural city's experimentation behaviors.

Importantly, since U-form departments are administered by the central ministries while M-form departments are administered by the regional political leaders, one should expect that the regional political leader's incentives mainly affect policy experimentations initiated by M-form ministries, rather than those initiated by U-form ministries. As we can see in Appendix Table A.6, this is indeed the case: the main results are almost entirely driven by policy experimentations initiated by M-form ministries. This 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 specific such that this confounding factor correlates with only M-form policy experimentations but not U-form policy experimentations.

Moreover, we find that when local politicians have strong career incentive and wants to take credit for the success of a policy experimentation, they also over-experiment in the intensive margin — by intentionally differentiating their policy experimentation from what their peers have already tried out previously with respect to the same policy, as shown in Appendix Table A.7. This could potentially be wasteful as the politician is abandoning policy solutions that have already been proven effective elsewhere.

7.3 Objectives of site selection beyond representativeness

The patterns on career incentives and political patronage documented above 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.

We first rule out a potential consideration: many speculate that the central government assigns policy experimentations to developed regions because the complexity of the reform requires high level of local governance and fiscal capacity to successfully implement the reform. Nonetheless, across several proxies for the complexity of the policies being experimented, we do *not* find evidence that more complex policies are systematically experimented in more positively selected localities (see Appendix Table A.8).

We next investigate a specific criterion concerning 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. Table 5, Panel C, presents the results among all policy experimentations. Column 1 controls for prefecture fixed effects, thus only exploiting within-region across-time variations in unrest occurrences; column 2 in addition controls for pre-experimentation characteristics such as local economic conditions; and column 3 in addition controls for time fixed effects. We find a robust pattern that prefectures that have experienced social and political unrest in the preceding period are significantly and substantially less likely to become experimentation sites, suggesting that unstable local environment could be a veto condition that preclude experimentations. Moreover, 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.9). 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.

The presence multi-dimensional objectives of the central government, as exemplified in the case of avoidance of local political unrest, could further push the experimentation sites selection away from being representative.[21]

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 C.

[21]There may exist several other objectives of the central government when choosing experimentation sites and conducting policy experimentations. For example, policy experimentations may be conducted in order for the central government to figure out implementation details, or to gather (potentially biased) evidence for internal political bargaining. We do not have direct evidence on these objectives.

8 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 four sets of results. First, the flat, flexible, and decentralized state organizational structure facilitates policy experimentation; as the ministries undergo reform from M- to U-form, the number of policy experimentations they administer declines. Second, among the policy experimentations that do take place, a substantial share deviates from representativeness in their implementation by oversampling more developed jurisdictions. Third, deviation from representativeness carries significant information cost, and biases policies to systematically favor more developed regions. Fourth, about half of the deviation from representativeness can be attributed to misaligned incentives between the central and local governments.

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.

References

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

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 and Esther Duflo, "The experimental approach to development economics," *Annu. Rev. Econ.*, 2009, *1* (1), 151–178.

_ , 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.

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.

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 Chinas 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

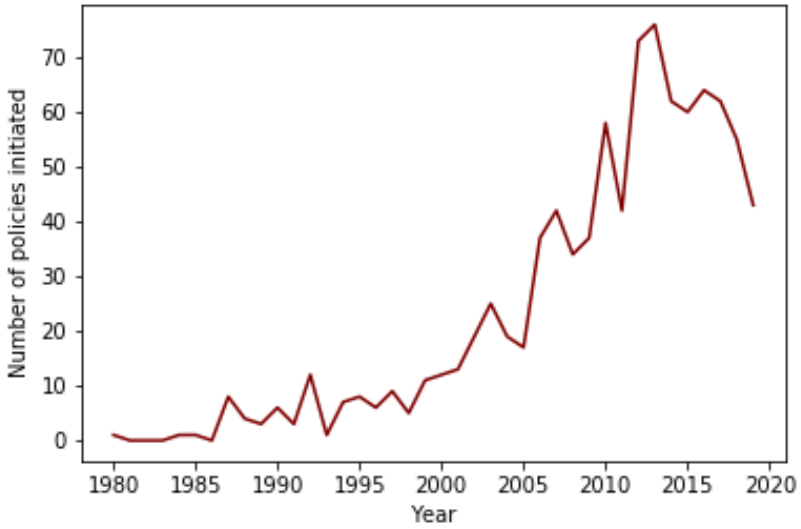


Figure 1: Count of policy experimentations by year. A policy with multiple rounds is counted only once at the year of its initial issuance.

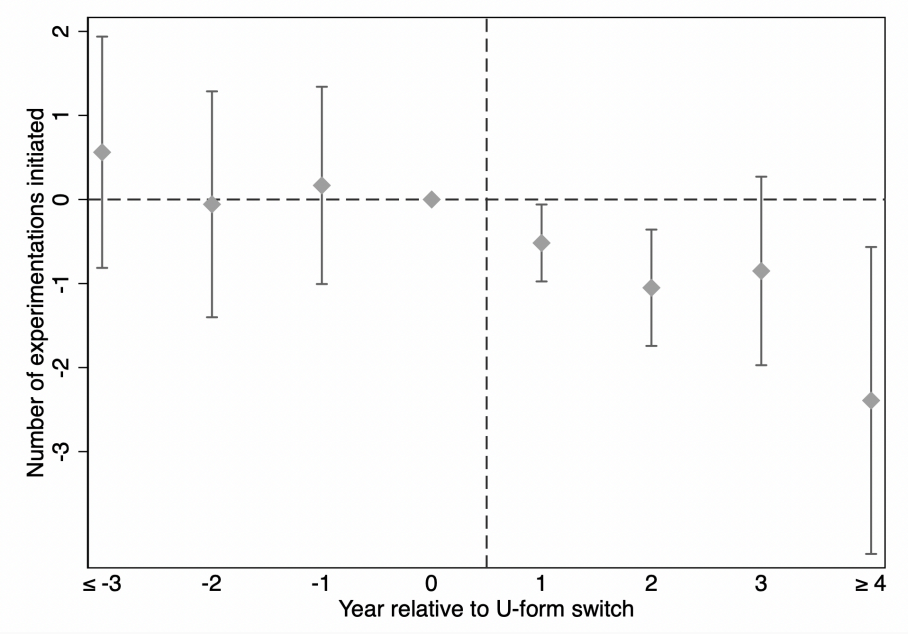


Figure 2: Count of policy experimentations initiated after transitioning into U-form. X-axis indicates the time relative to the reform. The point estimates and confidence intervals are computed from a standard event study design controlling for ministry fixed effect and calendar year fixed effect. Standard errors are clustered at the ministry level.

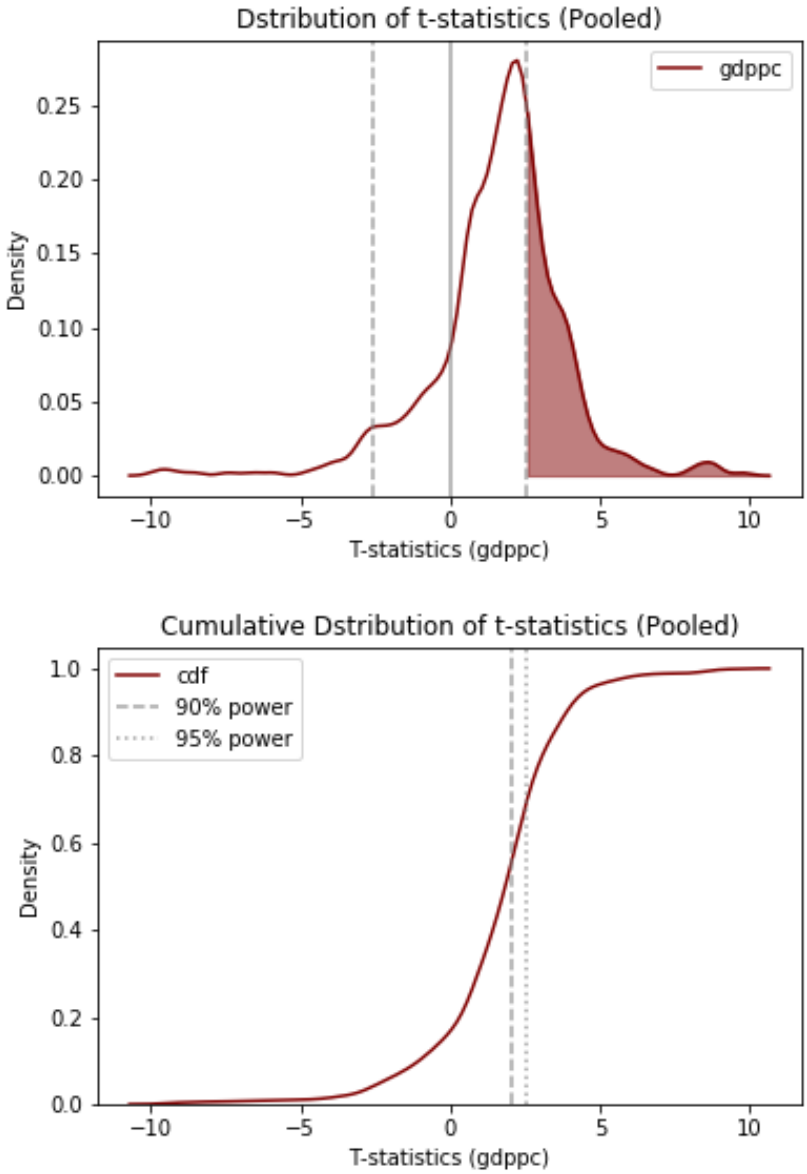


Figure 3: T-stat distribution from the representativeness test, based on GDP per capita. We plot the density in the upper panel and the cumulative distribution function in the lower panel. Within each test we compare the average pre-experimentation GDP per capita between those jurisdictions chosen as 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.

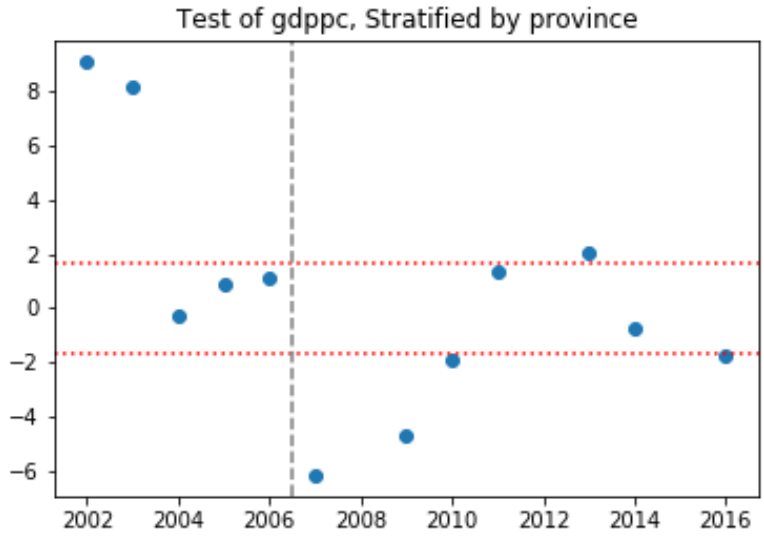


Figure 4: Local fiscal reform — representativeness test. We conduct stratified Fisher randomization tests with student-t statistics and provincial strata. Within each province, we view counties that engage in the experimentation for the first time as units of the treatment group, the rest as control. 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.

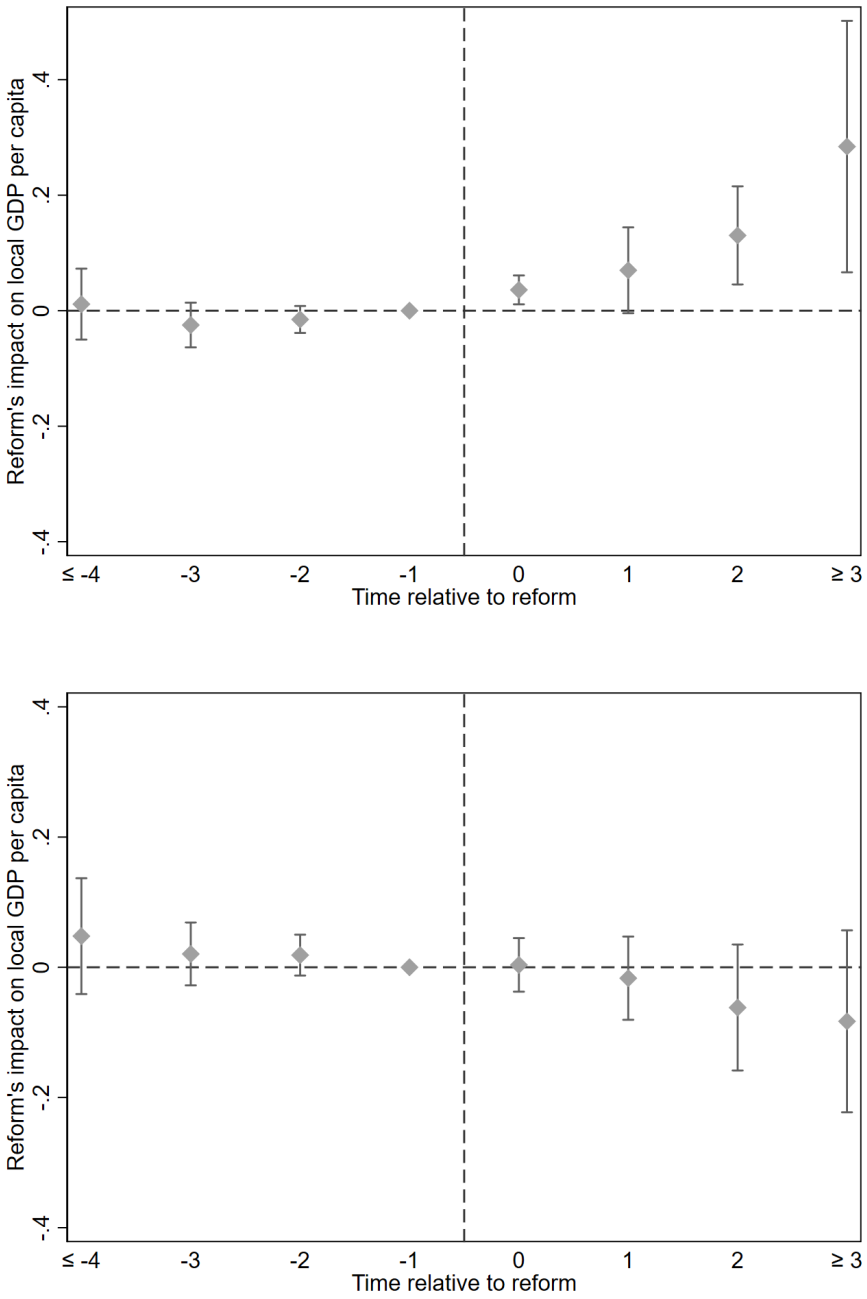


Figure 5: Local fiscal reform - treatment effect on local GDP per capita. Upper Panel plots the counties reformed before 2007 and Lower Panel plots those reformed after 2007.

Table 1: Summary statistics of policy experimentations

	# of exp.	# of rounds	# of sites	Roll-out rate
Panel A: full sample				
Full sample	854	2.83	15.69	0.235
Panel B: by policy domains				
Education	80	3.00	20.60	0.288
Agriculture	75	3.19	20.00	0.413
Market supervision	55	2.33	9.42	0.182
Finance & economics	49	2.43	5.78	0.082
Natural resources & environment	46	3.09	17.61	0.283
Business & commerce	41	4.27	21.73	0.390
Government finance & taxation	38	3.26	8.87	0.237
Population & health	33	2.73	23.27	0.394
Domestic affairs	33	2.48	25.61	0.424
Industrial information	33	2.64	21.64	0.212
Development & reform	32	2.72	18.84	0.219
General purpose	29	6.45	49.66	0.345
Transportation	24	1.79	10.46	0.083
Labor & personnel	21	2.14	17.10	0.095
Judiciary & supervision	10	2.00	10.10	0.100
Media	4	4.00	6.50	0.250

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	2.12	4.88	0.63
Panel A: by opt-in status			
Opt-in	2.48	5.62	0.60
Top-down	1.92	4.44	0.65
Panel B: by hierarchical level			
Province	-0.91	1.67	0.72
City	3.53	5.23	0.59
County	0.26	3.65	0.70

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 3.

Table 3: Similarity with experimentation sites and effects of policy roll-out

	Log growth rate of GDP per capita			
	All policies	All policies	Policies targeting econ-development	Policies with other objectives
	(1)	(2)	(3)	(4)
M-distance	-1.307***	-0.773***	-0.741**	-0.532
	(0.209)	(0.270)	(0.305)	(0.510)
Mean of DV	0.0935	0.0935	0.0845	0.110
County FE	NO	YES	YES	YES
Policy FE	YES	YES	YES	YES
# of obs.	96090	96090	61799	34287

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. Policy fixed effect, county fixed effect are subsequently controlled for across columns. We excluded those policies without explicit economic impacts in column 3, and report the estimated coefficients for the non-economic-policy subsample in column 4.

Table 4: Unrepresentative experimentations and subsequent inequality

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

Note: Robust standard errors are reported below the estimates. A policy is considered *representative* if we cannot reject the null of representativeness at 10% level, in terms of pre-experimentation GDP per capita. In column 1, we regress the annual inter-county Gini index on the number of unrepresentative policies after their nationwide roll-out. In columns 3, we horse-race both types of policies. In columns 2 and 4, we restrict the sample to important policies mentioned in the *Five Year Plans* of the Standing Committee.

Table 5: Political incentives and policy experimentations

	Engage in experimentations		
	(1)	(2)	(3)
Panel A: Political patronage			
Connected to minister	0.059***	0.037**	0.037**
	(0.019)	(0.014)	(0.016)
Controls	NO	NO	YES
Time FE	NO	YES	YES
Ministry by province FE	YES	YES	YES
Panel B: Career incentive			
Career incentive	0.397**	0.534**	0.469**
	(0.200)	(0.220)	(0.207)
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 event	-0.003***	-0.002**	-0.002***
	(0.001)	(0.001)	(0.001)
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. For regression results in other sub-samples (M-form/U-form, top-down assignment/voluntary opt-in), please refer to Appendix Tables A.5, A.6, and A.9.

APPENDIX

A Optimal experimental design simulations

We simulate the experiment described in Section 6.1 to benchmark the real-world experimental design against the theoretically optimal one under two models.

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

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, subjective utility maximization is the case where $\lambda = 1$.

We simulate the optimal experimental design with the following procedure:

1. We first compute the vector of treatment effects p for each county that receives treatment, using a difference-in-difference specification with controls for pre-experiment GDP and province fixed effects. There are 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 2010 counties.

   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 2010 bins corresponding to the 2010 counties.

2. Next, we construct the space of priors H. Each prior consists of 2010 expected treatment effects (one per county) $prior_{p,c} \in h_p \in H$, with the following the data generation process:

   $$prior_{p,c} = \beta_c + \gamma_p + \eta_{p,c} \qquad \gamma \sim U[-2\bar{\beta}, 2\bar{\beta}], \eta \sim U[-\beta_{max}, \beta_{max}]$$

   where p indexes a particular 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 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 prior's treatment effect is the true treatment effect. We

construct 1000 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 597 counties to be treated in the experiment. 1000 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 potential solutions to the DP.

4. Once the priors and potential DP solutions have been construct, 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 compare it to that found in reality.

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

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

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 A.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.

B Constructing index for career incentive

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 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:

$$\hat{y}_{pt} = \Phi^{-1}\left\{\hat{\alpha} \cdot startage_{pt} + \hat{\beta} \cdot level_{pt} + \hat{\gamma} \cdot startage_{pt} \times level_{pt}\right\}. \tag{9}$$

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.10 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).

C 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 (from Section 7.2.1), career incentive (from Section 7.2.2), and political unrest (from Section 7.3). For the baseline, we estimate the following econometric model using policy-prefecture level data:

$$y_{cp} = \alpha \cdot lgdppc_{cp} + Distortions_{cp}^{'}\beta + \gamma_p + \epsilon_{cp}. \tag{10}$$

Appendix Table A.11 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 10, 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.16, Panel B, shows the distribution of t statistics of the representative test, as described in Section 5.1, when we assert a non-stochastic version of treatment assignment mechanism. In this setting, those prefectural units with the top k propensity score get chosen as experimentation spots, where k corresponds to the number of sites chosen for each policy at status quo. Compared with our baseline specification shown in Appendix Figure A.16, 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.16, Panel C. This specification is most similar, in general ideas, to the regression presented in Table A.11, confirming the idea that all distortion factors we identified explain almost half of the selection bias of policy experimentation.

D Additional figures and tables

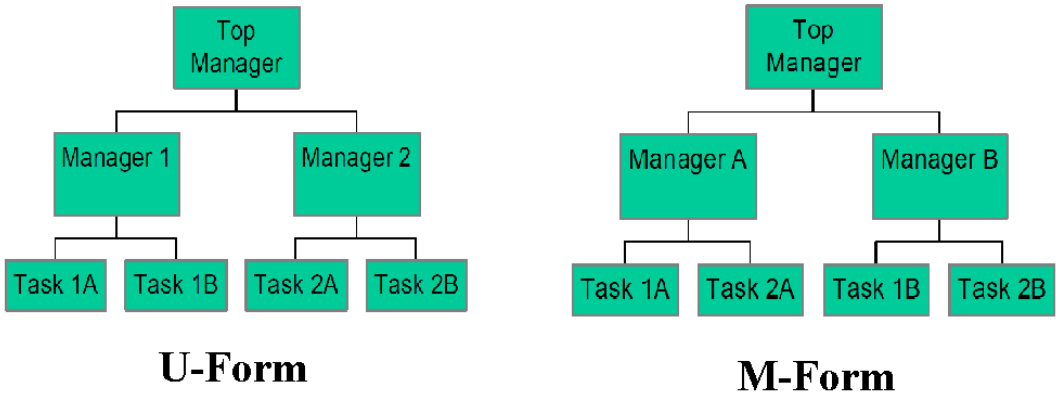


Figure A.1: Reproduced from Qian et al. (2006). Illustration of a shift from M form to U form. The top manager (ministers at central government branches in our case) have more administrative and personnel authority on its branches.

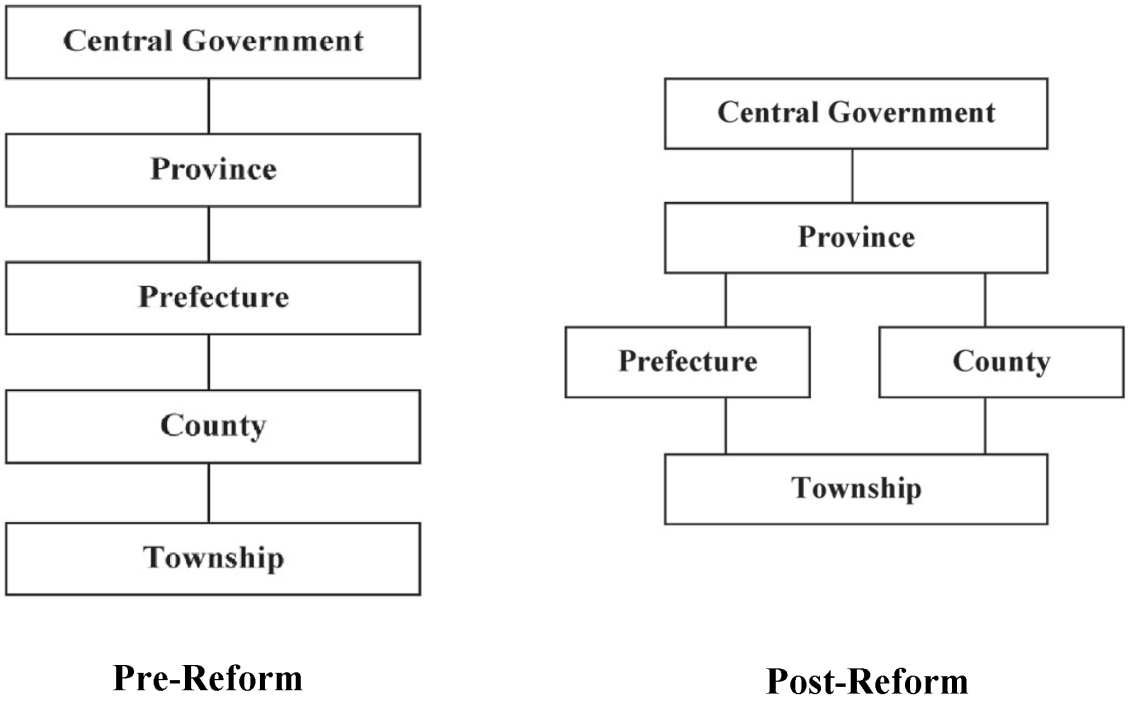


Figure A.2: Reproduced from Li et al. (2016). Illustration of local fiscal reform. After the reform, the provincial government could directly manage some of its counties, bypassing the prefectural cities, which grants county governments with more fiscal autonomy.

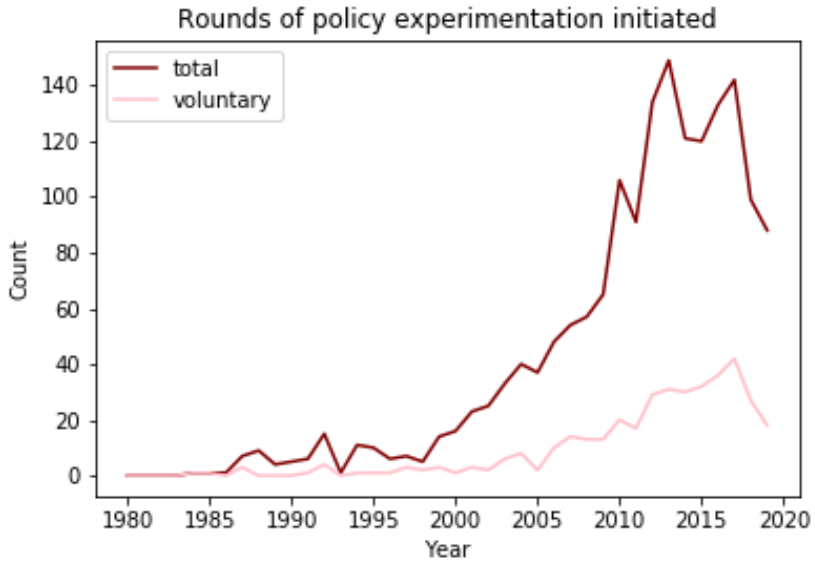


Figure A.3: Count of rounds of policy experimentation, by year. We group the documents issued within 30 days together as one round and define a policy round as voluntary if at least one document in this round involves voluntary opt-in.

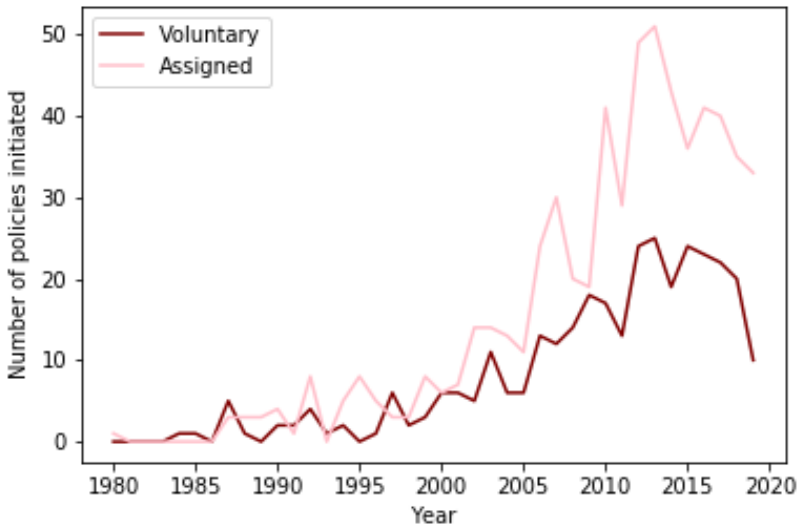
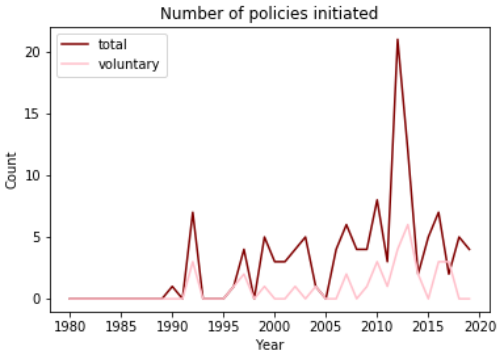
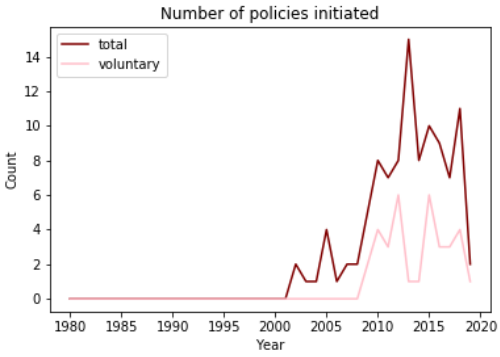


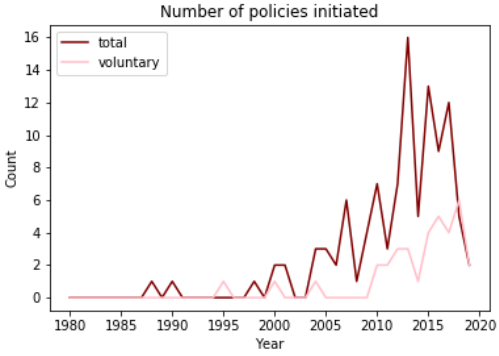
Figure A.4: Count of assigned/voluntary policy experimentation, by year. We define a policy experimentation as voluntary if at least one round of this policy involves voluntary opt-in; otherwise it is a policy totally assigned by the central government.



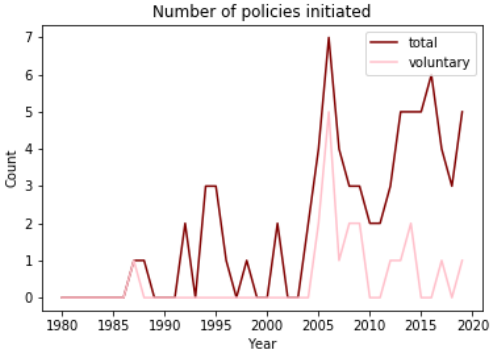
Education



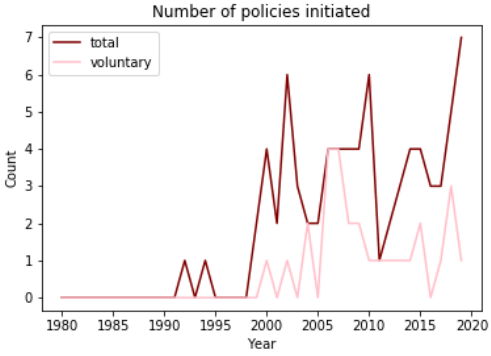
Agriculture



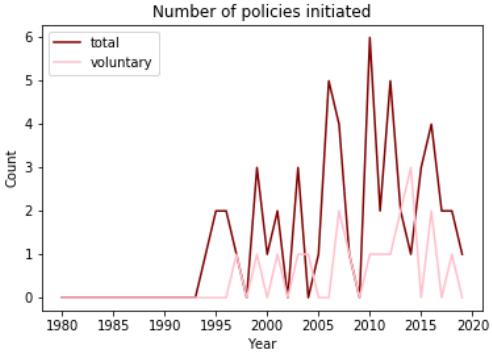
Market supervision



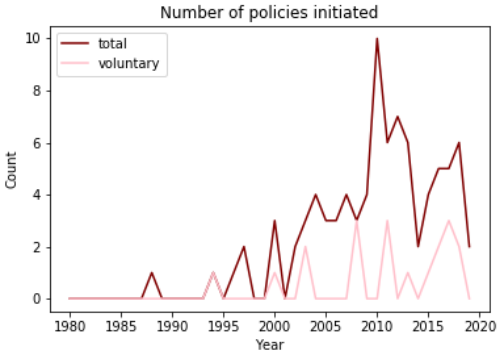
Finance & economics



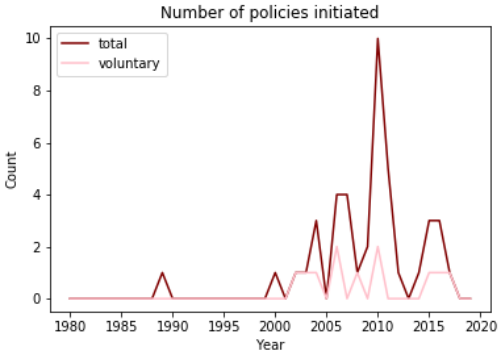
Natural resources & environment



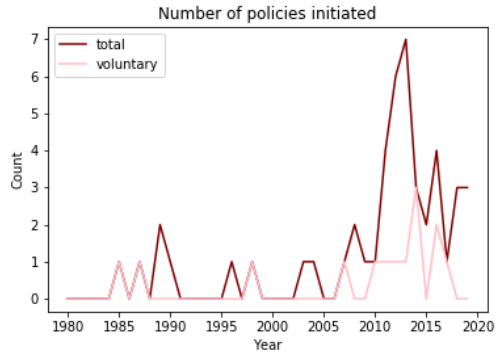
Business & commerce



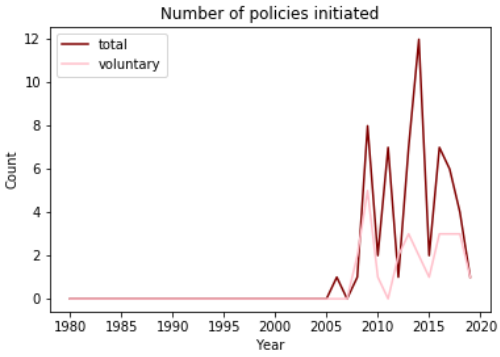
Government finance & taxation



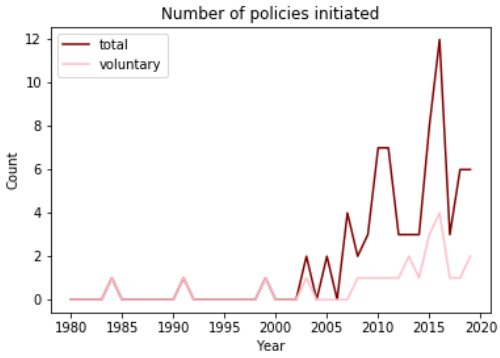
Population & health



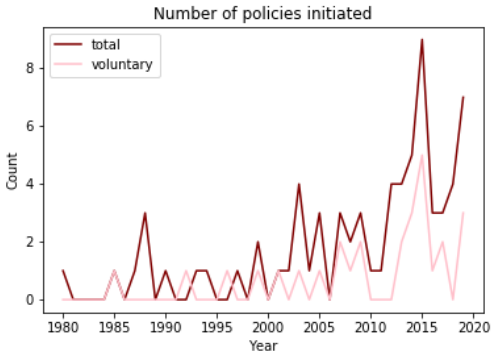
Domestic affairs



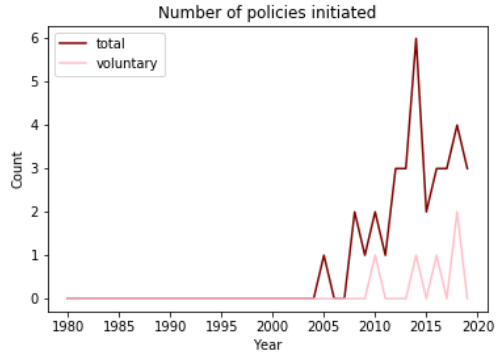
Industrial information



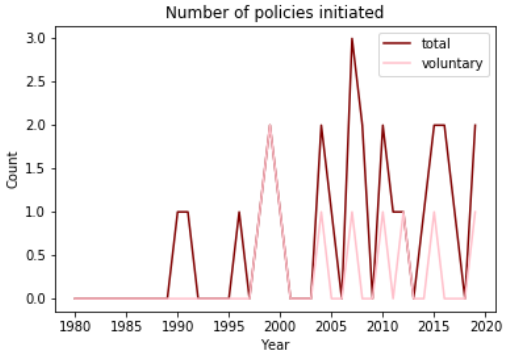
Development & reform



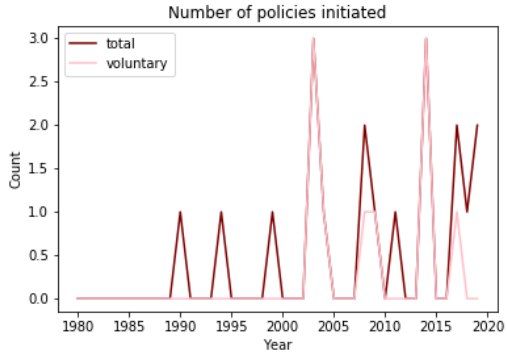
General purpose



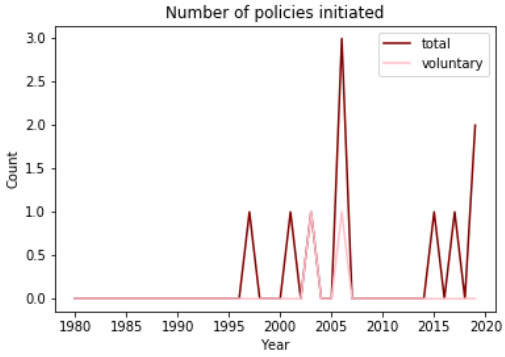
Transportation



Labor & personnel



Judiciary & supervision



Media

Figure A.5: Count of policy experimentation, by ministerial function.

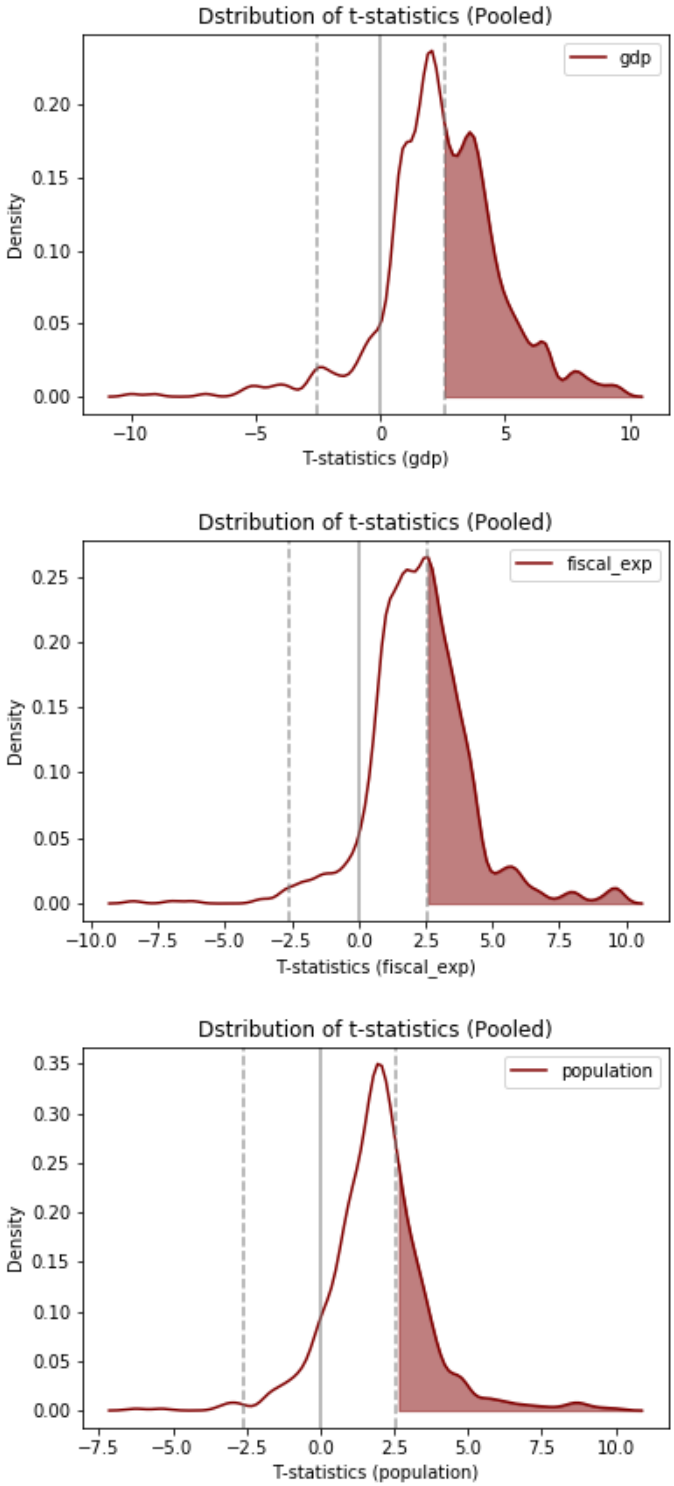


Figure A.6: Representativeness tests' t-statistics distribution, using other features including fiscal expense, GDP and population, with the same test procedures as Figure 3.

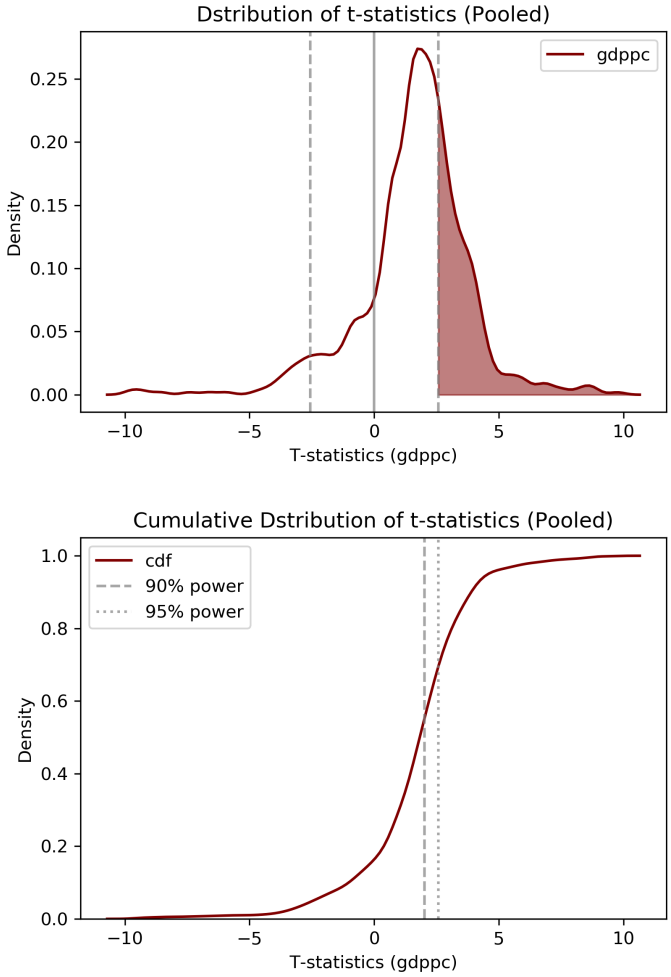


Figure A.7: Representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. One site policies are pooled chronologically by clusters n=5.

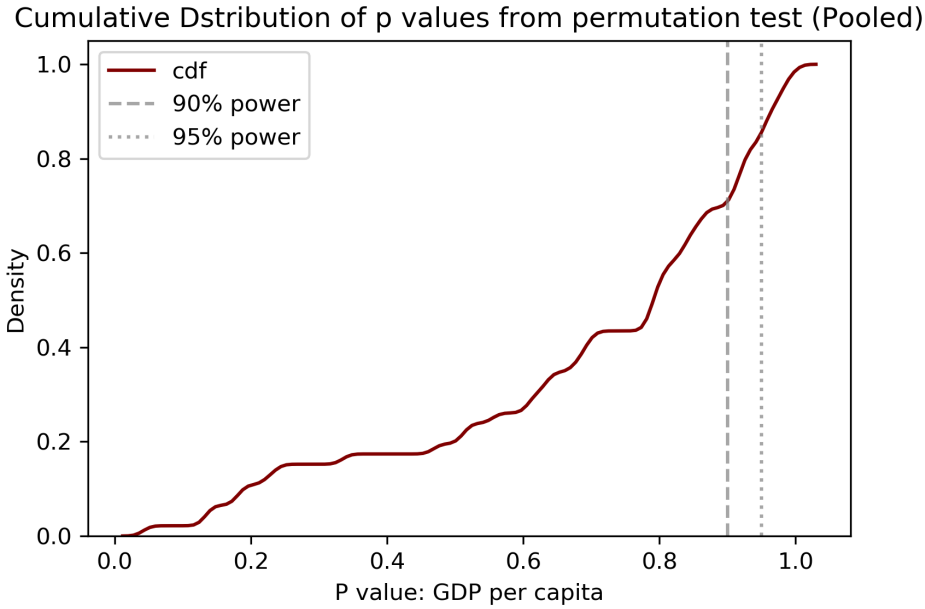


Figure A.8: Cumulative distribution of p values from permutation tests in representativeness tests. Each realized student-t statistic is compared with 5,000 permuted t values to calculate the p statistic. In small samples, permutation tests are more conservative than standard t tests.

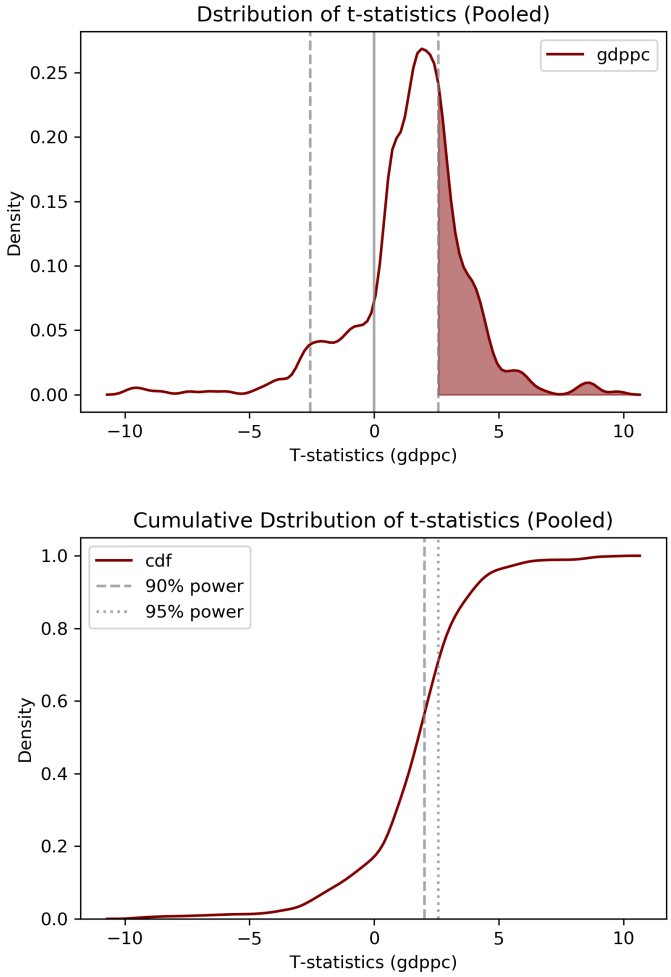


Figure A.9: Representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. One site policies are pooled by bootstrapping with replacement for 166 times, without specifying the existence of any certain policy, and concatenated to the multi-site sample.

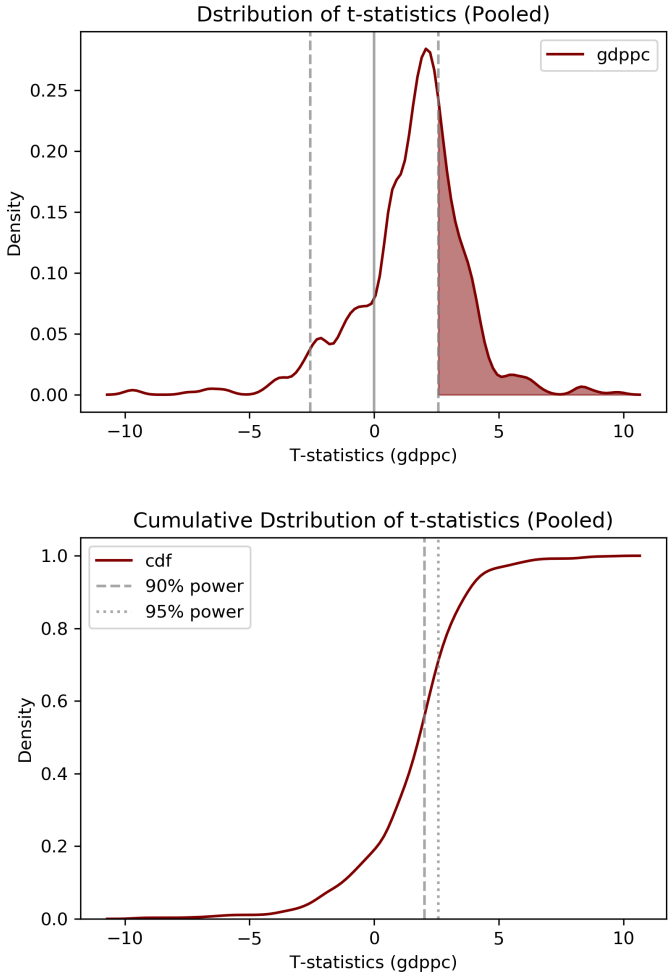


Figure A.10: Representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. Only policies in early rounds are considered.

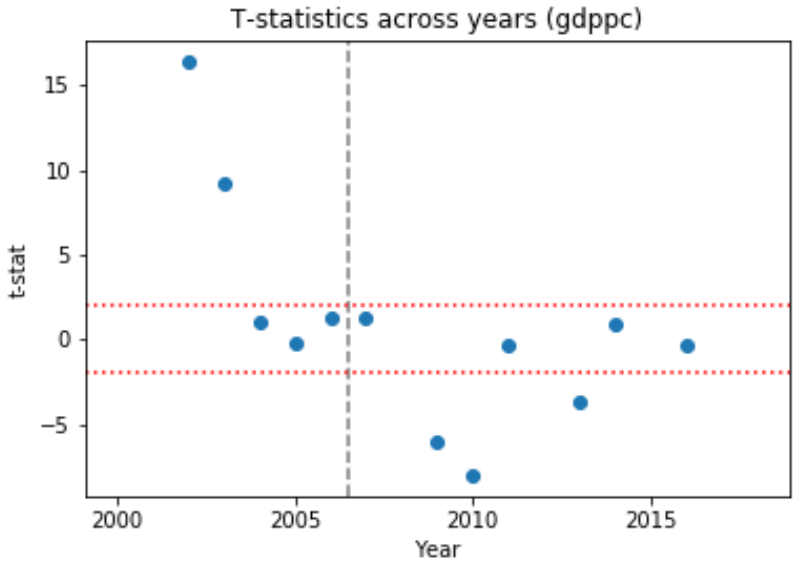


Figure A.11: Local fiscal reform — representativeness test, without province FE. Same test procedure as Figure 3. No provincial clusters are taken into account. Compared with the last figure, the underlying assumption of this test is that the choice of treated units is decided entirely by the central government, whereas in Figure 4 we assume that the decision is made at provincial government level.

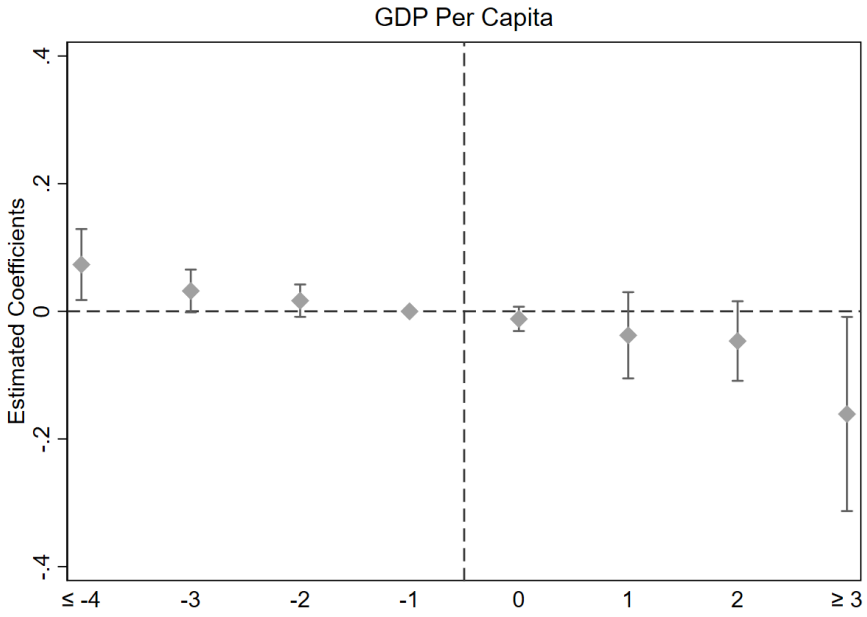


Figure A.12: Local fiscal reform - poor counties experimented before 2008. The pattern demonstrated here is similar to that in Figure 5, Panel B.

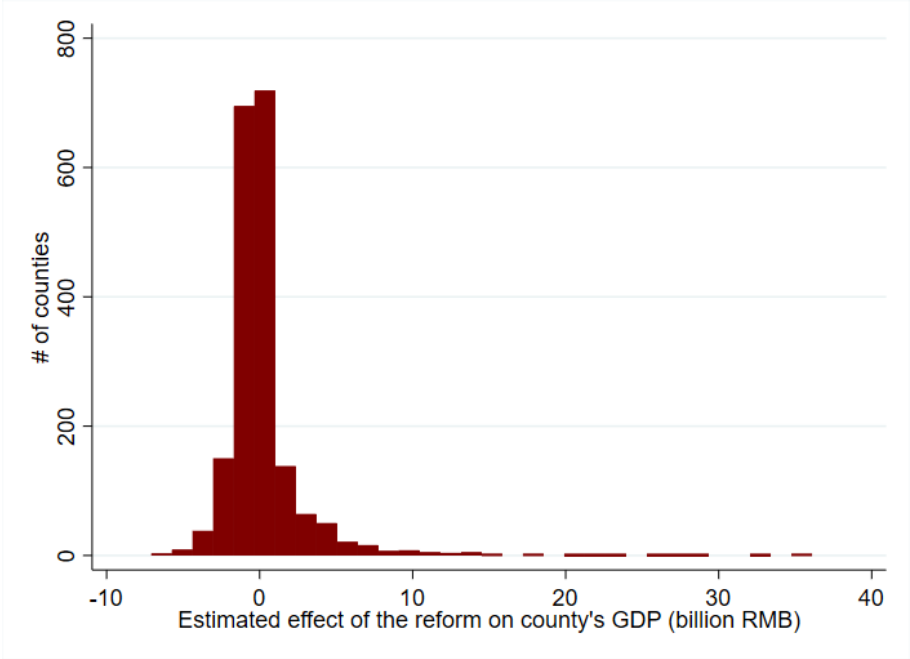


Figure A.13: Simulated treatment effects across country. We extrapolate the estimated treatment effect to all counties nationwide and obtain a distribution of reform effect on county's GDP.

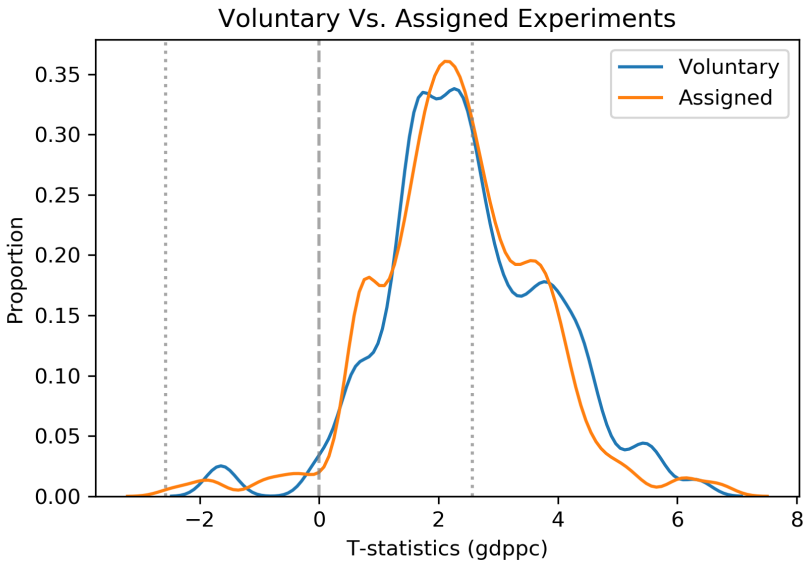


Figure A.14: Representativeness tests' t-statistics distribution using GDP per capita, distinguished by voluntary and assigned experimentations, with the same test procedures as Figure 3.

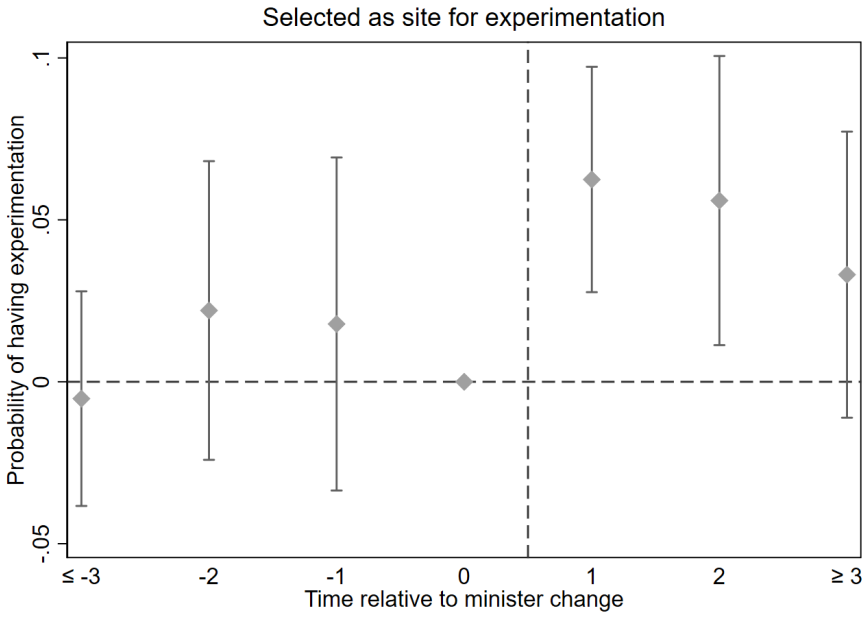


Figure A.15: Probability of being selected as site for experimentation after being related to a ministry due to central minister turnovers.

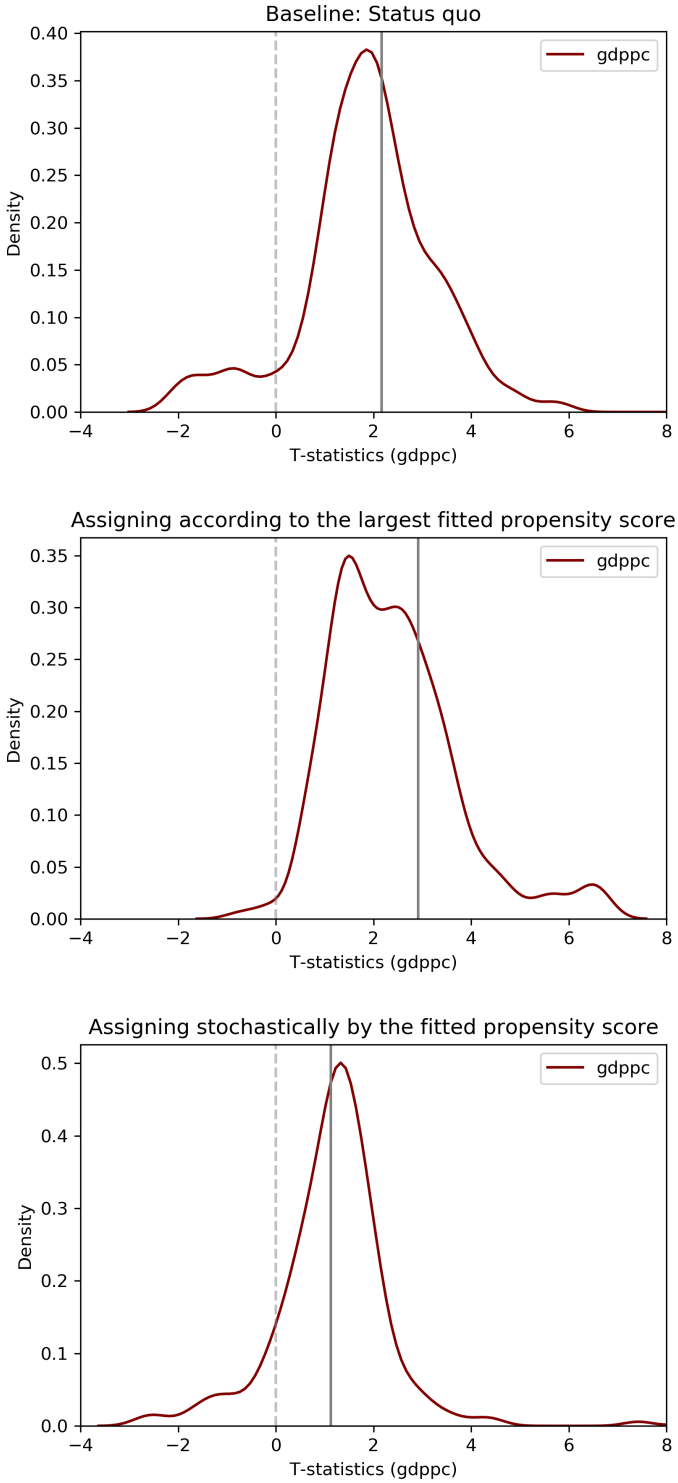


Figure A.16: Representativeness tests' t-statistics distribution, using GDP per capita with the same test procedures as Figure 3. To make sure we're making reasonable comparisons, we adjust the baseline Panel A so that it only includes the experimentations targeting prefectural cities, with 4 municipal cities excluded from observation. Panel B shows the simulated results when sites for experimentation are assigned to the prefectural units with the largest fitted propensity score. Panal C imposes a milder assumption where we assert a stochasitic version. Extreme values are trimmed. The grey vertical line indicates the mean of t-stats.

Table A.1: 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.2: 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.3: 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.4: Representativeness of experimentation sites selection and policy's national roll-out

	National roll-out							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Representative	-0.217***	-0.212***	-0.233***	-0.223***				
	(0.038)	(0.038)	(0.053)	(0.047)				
T-stats					0.038***	0.037***	0.037***	0.038***
					(0.007)	(0.007)	(0.012)	(0.010)
Voluntary		0.094**	0.123**	0.122**		0.088**	0.106*	0.106*
		(0.038)	(0.052)	(0.050)		(0.039)	(0.057)	(0.056)
Important policies			0.102*	0.102			0.127**	0.126
			(0.055)	(0.069)			(0.056)	(0.074)
Mean of DV	0.339	0.339	0.339	0.339	0.339	0.339	0.339	0.339
Controls for implementation levels	YES	YES	YES	YES	YES	YES	YES	YES
Ministry FE	NO	NO	YES	YES	NO	NO	YES	YES
Year FE	NO	NO	NO	YES	NO	NO	NO	YES
# of obs.	486	486	355	353	486	486	355	353

Note: The standard errors clustered by ministries are reported below the estimates. In columns 1-4, we regress the rollout dummy on an indicator for representativeness, whereas in columns 5-8, we regress it on T-stats, whose construction is described in Section 5.1, to allow for more variation. One-site experimentations are excluded from the regression.

Table A.5: Political patronage and engagement in experimentations

	Engaged in experimentations		
	(1)	(2)	(3)
Panel A: All experimentations			
Connected to minister	0.059***	0.037**	0.037**
	(0.019)	(0.014)	(0.016)
Panel B.1: Experimentations initiated by M-form ministry			
Connected to minister	0.057**	0.021	0.025
	(0.021)	(0.016)	(0.017)
Panel B.2: Experimentations initiated by U-form ministry			
Connected to minister	0.043	0.023	-0.004
	(0.043)	(0.041)	(0.047)
Panel C.1: Experimentations with top-down assignments			
Connected to minister	0.057***	0.039***	0.040***
	(0.018)	(0.014)	(0.014)
Panel C.2: Experimentations with voluntary sign-ups			
Connected to minister	0.018	0.009	0.011
	(0.016)	(0.014)	(0.017)
Mean of DV	0.134	0.134	0.134
Controls	NO	NO	YES
Time FE	NO	YES	YES
Ministry by province FE	YES	YES	YES
# of obs.	57784	57784	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 counties' 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.6: 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 B.

In 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.7: Political career incentives and differentiation in experimentations

	Similarity in experimentation proposal		
	(1)	(2)	(3)
Panel A: All experimentations			
Career incentive	-0.065***	-0.064***	-0.065***
	(0.021)	(0.021)	(0.022)
Panel B.1: Experimentations initiated by M-form ministry			
Career incentive	-0.070***	-0.070***	-0.071***
	(0.022)	(0.022)	(0.023)
Panel B.2: Experimentations initiated by U-form ministry			
Career incentive	0.022	0.008	0.026
	(0.053)	(0.054)	(0.051)
Panel C.1: Experimentations with voluntary sign-ups			
Career incentive	-0.095***	-0.095***	-0.090**
	(0.036)	(0.036)	(0.037)
Panel C.2: Experimentations with top-down assignments			
Career incentive	-0.041	-0.038	-0.042
	(0.025)	(0.026)	(0.027)
Mean of DV	0.922	0.922	0.922
Round FE	NO	NO	YES
Time FE	NO	YES	YES
Policy FE	YES	YES	YES
# of obs.	3145	3145	3145

Note: The standard errors clustered by policies are reported below the estimates. Career incentive index is constructed following the same procedures as explained in Appendix B
An index representing similarity in experimentation proposal is constructed from the matrices of pairwise text similarity between policy documents on the same topic. Each entry of the matrix is calculated via Paddlehub, an open source NLP model by Baidu. The final index is retrieved by taking row-wise maximum of the lower-triangular matrix. This is equivalent to comparing the similarity between each document and those issued previously on the same policy, and finding the minimum deviation from its peers. The result is robust to outliers or irrelevant documents mis-identified since the operation of taking maximum narrows our focus to one's closest counterpart.

Table A.8: Complexity of policy experimentations and representativeness in site selection

	Representative indicator				t-statistics			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
# of departments involved	-0.079	-0.062			-1.202***	-1.518***		
	(0.063)	(0.064)			(0.415)	(0.496)		
Average length of documents			0.015	0.014			-0.511	-0.491
			(0.023)	(0.023)			(0.412)	(0.406)
Mean of DV	0.597	0.597	0.597	0.597	3.867	3.867	3.867	3.867
Controls	NO	YES	NO	YES	NO	YES	NO	YES
# of obs.	534	534	534	534	478	478	478	478

Note: Here we focus on a subset of policies that experimented at the prefectural level. Variables of interest are both in logarithm. In columns 1 to 4 we regress on the dummy variable for representativeness, and in column 5 to 8, the dependent variable is the t-statistics from representativeness tests.
The robust standard errors are reported below the estimates. Control variables include whether the experimentation sites are selected from top-down assignment or voluntary sign-up, and whether the ministry in charge is M-form or U-form.

Table A.9: 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.10: Predicting politicians' career incentives

	Promotion			
	OLS (1)	OLS (2)	Probit (3)	Probit (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.11: Political incentives and engagement in experimentations

	Engaged in experimentations			
	(1)	(2)	(3)	(4)
GDP per capita	0.019***	0.009***	0.037***	0.021***
	(0.001)	(0.001)	(0.002)	(0.001)
Mean of DV	0.016	0.016	0.016	0.016
Controls for political distortion	NO	YES	NO	YES
Policy FE	NO	NO	YES	YES
# of obs.	100420	97693	100420	97693

Note: The robust standard errors for (columns 1 & 2), and standard errors clustered at policies (column 3 & 4) are reported below the estimates. The controls for political distortion include the career incentives of prefecture party leader, as described in Appendix B, 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 7.2.1). 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.