Curbing Bureaucratic Information Manipulation: Evidence from a Statistical Reform in China*

Yongwei Nian

December 17, 2024

Abstract

Bureaucrats are often incentivized to manipulate information, which may have real consequences. Leveraging China's 2009 reform punishing economic data manipulation and counties' quasi-random reform exposure, this paper provides rich causal evidence showing that the reform led to: (1) a decrease in GDP growth manipulation amounting to 5% of reported GDP growth, driven by a *reputational discipline* effect; (2) an increase in local officials' development effort manifested in both policy changes and downstream impacts, consistent with an *effort reallocation* effect. These results thus highlight the far-reaching costs of such manipulation and the welfare implications of curbing it.

*School of Economics, Peking University. Email: yongweinian@gmail.com. Website: www.yongweinian.com. I thank Stefano Fiorin, Guido Tabellini, and Paolo Pinotti for their invaluable advice. I also thank Jan Bakker, Jean-Marie Baland, Lee Benham, Morten Bennedsen, Martina Björkman Nyqvist, Luca Braghieri, Alexia Delfino, Erika Deserranno, Sarah Eichmeyer, James Fenske, Shiqi Guo, Rafael Jimenez-Duran, Anke Kessler, Zanhui Liu, Jiaxuan Lu, Marco Ottaviani, Eleonora Patacchini, Jaime Marques Pereira, Carlo Schwarz, Liyang Sun, Silvia Vannutelli, Leonard Wantchekon, Liam Wren-Lewis, and all participants at Bocconi Political Economy Breakfast, Bocconi DLPE seminar, Bocconi Applied Micro seminar, EUDN Development Economics Workshop 2023, Asian Meeting of the Econometric Society 2023, EEA-ESEM 2023, Society for Institutional & Organizational Economics Annual Meeting 2023, and Oxford Development Economics Workshop 2023 for their useful discussions and suggestions.

1 Introduction

Bureaucracies are pivotal to economic development, a view that dates back at least to Max Weber and is further stressed by a growing body of economics literature (Besley et al., 2021). However, lower-level bureaucrats may have incentives to manipulate information, especially when it is instrumental to their career advancement but imperfectly observed, thereby undermining bureaucratic effectiveness. One example is the overreporting of economic statistics by local officials in China, where their career advancement is closely tied to economic performance (Wallace, 2016; Xiong, 2018; Lyu et al., 2018). Examples outside China and the domain of economic statistics also abound (Bossuroy, Delavallade and Pons, 2019; Martinez, 2022; Kofanov et al., 2023). Such manipulation may not only create information distortion in the bureaucracy, but may also divert bureaucrats' effort from their designated missions from a multitasking perspective (Holmstrom and Milgrom, 1991). Despite its prevalence and potential costs, there is scant evidence on how to effectively curtail such manipulation, and even less is known about the economic benefits of potential interventions.

This paper sheds some of the first light on these issues leveraging China's 2009 reform aimed at combating economic data manipulation by local officials. The reform, initiated by the National Bureau of Statistics (NBS) and other central authorities, increased punishment for manipulation nationwide. However, the detection mainly relied on preexisting survey teams in 40% of counties, creating cross-sectional variations in counties' exposure to the reform. When manipulation was detected, the survey teams could directly report to the central authorities, typically leading to reputational warnings or demerits as punishment. Notably, these survey teams were deployed in 2005 and initially tasked with economic surveys, making their deployment largely unrelated to local data manipulation. This unique setup reduces concerns of positive selection common in government interventions (Al-Ubaydli, List and Suskind, 2019; Wang and Yang, 2021). In addition, the teams operated independently from local political influence under the leadership of the NBS, a key feature for effective monitoring (Kofman and Lawarrée, 1993; Olken, 2007; Vannutelli, 2022).

To identify the effects of the reform, I assemble a county-level dataset from 2005 to 2018 and employ a difference-in-differences design comparing counties with these survey teams (treatment) to counties without these survey teams (control) before and after the launch of the reform in 2009. I focus on the manipulation of GDP growth given its central role in dictating local officials' career advancement (Li and Zhou, 2005). I find that the reform significantly decreased GDP growth manipulation, which is measured as the discrepancy between GDP growth reported by local statistics bureaus and nighttime light intensity growth inferred from satellite observations, following the seminal framework of Henderson, Storeygard and Weil (2012). In terms of economic magnitude, in the most stringent specification with a set of baseline demographic, economic, and geographic controls interacted with the post-reform dummy, I find a 0.58 percentage points drop in GDP growth manipulation in treatment counties relative to control counties, which amounts to 5.3% of the mean of reported GDP growth. I further find no pre-trends in manipulation between treatment and control counties using an event study specification, which is consistent with the pre-deployment nature of the survey teams.

I conduct various tests to ensure the robustness of the baseline findings. First, in the spirit of Martinez (2022), I show that the estimates on GDP growth manipulation remain virtually unchanged when I allow the mapping between economic growth and light growth to be nonlinear or depend on various temporal and spatial characteristics. Second, I show that the results are robust to accounting for potential imbalance between treatment and control counties, by achieving covariate balance through entropy balancing Hainmueller (2012) or coarsened exact matching (Iacus, King and Porro, 2012), among other conventional methods. Third, I conduct a placebo test by examining the dynamic effect around 2005 when the survey teams were deployed but had not conducted any disciplining actions. This test helps to further rule out pre-trends and alleviate additional concerns that these teams per se may affect my outcomes, even in the absence of the reform in 2009. Reassuringly, I find no diverging trends between treatment and control counties until 2009.

To further address concerns about unobserved heterogeneity, I leverage a unique institutional feature to conduct an instrumental variable estimation. By checking local gazetteers, I find that most of these survey teams launched in 2005 were restructured from earlier rural survey teams set up in 1984. As I describe later, these earlier teams were mainly used to collect information on agricultural output and their assignment at that time was done through a systematic random sampling of counties within provinces. This random assignment hence forms a valid instrument for counties with survey teams deployed in 2005. The validity of this instrument is further supported by balance tests showing that counties with these earlier rural survey teams were no different from other counties on a battery of baseline county characteristics. In addition, counties with these earlier rural survey teams were in parallel with other counties in terms of GDP growth manipulation until 2009. The difference-in-differences estimates using this instrument have no substantial changes, suggesting that the baseline findings are unlikely to be driven by unobserved heterogeneity between treatment and control counties.

I investigate several potential mechanisms that could account for the drop in GDP growth manipulation, following the predictions from a conceptual framework. Specifically, the reform could decrease GDP growth manipulation through both a *discipline* effect and a *selection* effect. The discipline effect refers to behavioral changes within local officials' terms due to reputational and promotional costs, while the selection effect stems from the removal of those involved in manipulation or the appointment of more competent successors. To distinguish between these two mechanisms, I focus on a sample in which there was no change of local officials in the treatment counties. The estimates are similar to those in the full sample. I also show that the reform had no effects on personnel turnovers and official characteristics. Hence, the findings can only be explained by the reform generating a discipline effect on local officials. This discipline effect could be further decomposed into a reputational discipline effect and a promotional discipline effect. To distinguish between these two mechanisms, I show that the decrease in GDP growth manipulation is no different for local officials with greater promotion incentives, which are measured using either age restrictions on promotion or estimated ex ante likelihood of promotion. Hence, the findings so far are most consistent with a reputational discipline effect caused by the exposure of manipulation within the bureaucracy upon detection. This result resonates with a growing literature showing both theoretically and empirically the role of reputational concerns in shaping truth-telling (Benabou and Tirole, 2011; Gneezy, Kajackaite and Sobel, 2018; Abeler, Nosenzo and Raymond, 2019). I also rule out some alternative explanations. First, I show that the reduction in manipulation was not driven by potential soft information acquired by the survey teams, which may dampen the role of GDP growth in performance evaluation. Second, I show that the reduction in manipulation was not driven by improvements in local statistical capacity, which may close the gap between reported GDP growth and light growth. Finally, I show that the findings did not capture the effects of other concurrent reforms that may also strengthen the monitoring of local officials, such as the anti-corruption inspections.

I then test the effect on local officials' effort in developing the economy. Conceptually, local officials should allocate more effort to developing the economy given the relatively higher cost of GDP growth manipulation after the reform. Along this line, I first show through a textual analysis of county government work reports that government policies shifted in directions conducive to economic growth after the reform. In particular, I find that local officials put more emphasis on business attraction and market reform, which are critical to China's recent economic success (Xu, 2011). To alleviate concerns about cheap talk in these reports and shed light on whether these policy shifts translated into any real impacts, I further examine two downstream economic outcomes closely related to local officials' effort. The first is bank credit, over which local officials have substantial discretion (Ru, 2018; Cong et al., 2019). I show that the reform generated positive impacts on various measures of bank credit, and such effects were more pronounced for loans to small firms and credit from banks controlled by local governments. The second is firm entry, which is subject to lengthy bureaucratic procedures in China (WorldBank, 2008; Jia, Lan and Padró i Miquel, 2021). I also find that the reform boosted firm entry, especially for those with higher productivity. Finally, I investigate citizens' trust in local officials and evaluation of local government performance, two common measures of the performance of government officials (Bertrand et al., 2020; Martinez-Bravo et al., 2022). I find that citizens' attitudes towards local officials improved after the reform; as a placebo, their trust in most people or evaluation of others' health had no changes. Overall, these results are consistent with local officials exerting more effort in economic development after the reform. I also rule out several alternative explanations. First, I show that local officials' differential performance did not result from varying fiscal transfers from the central government, which may have similar development effects. Second, I show that better policies adopted by local officials was not driven by the reform facilitating policy diffusion among counties or from upper-level governments. Third, I show that local officials did not switch to other short-termist behaviors that may boost the economy in the short run, such as overleverage. Finally, using both corruption convictions and perceived corruption by citizens, I show that their improved performance was not due to a decrease in corruption due to perceived higher corruption cost after the reform.

This paper relates to several strands of literature. First, it contributes to the nascent literature documenting the manipulation of statistical data by career-minded government officials across various domains and political regimes. A common form of such manipulation is the inflation of economic statistics in authoritarian countries (Wallace, 2016; Lyu et al., 2018; Xiong, 2018; Martinez, 2022). Additionally, such manipulation exists in other settings, such as the underreporting of pandemic statistics in Russia (Kofanov et al., 2023), the falsification of regional electricity consumption bills in India (Mahadevan, 2023), and the manipulation of air pollution data in China (Ghanem and Zhang, 2014; Greenstone et al., 2022). Despite the pervasiveness, there is surprisingly little evidence on how to curb such manipulation. One possible reason is the underestimation of the potential costs. In certain cases, the costs are obvious, such as profit losses by utility providers (Mahadevan, 2023) or the underinvestment of defensive goods against pollution by citizens (Greenstone et al., 2022). In the domain of economic data manipulation, the costs are not obvious. My paper thus advances this literature in two ways. First, to my best, this is the first to estimate the causal impacts of a large-scale intervention on the exaggeration of GDP statistics by local officials, a typical form of such manipulation. Second, the rich micro-level data allows me to further estimate the downstream impacts of this intervention. By highlighting the real consequences of such manipulation, this paper thus has implications for the design of relevant interventions targeting economic data manipulation, which are still scarce.

Second, this paper also adds to a growing literature on disciplining local officials. In the spirit of Becker and Stigler (1974), existing research shows that the combination of monitoring and punishment could reduce unaccounted expenditures in Indonesian villages (Olken, 2007), curb the misuse of federal funds in Brazilian municipalities (Avis, Ferraz and Finan, 2018), and enhance fiscal performance in Italian local governments (Vannutelli, 2022). Underpinning the success of such practices are two noteworthy features. On the monitoring side, independent audit in a top-down fashion is critical. Lack of independence may lead to collusion between monitors and local officials (Kofman and Lawarrée, 1993). My study features such a setting where the survey teams were deployed from the central and insulated from local interference. On the punishment side, extant research highlights the role of electoral and judicial punishment. My study differs from this, however, by showing that the findings are most consistent with a reputational discipline effect. In this vein, my findings join a small literature showing the role of reputational concerns in shaping truthtelling (Benabou and Tirole, 2011; Gneezy, Kajackaite and Sobel, 2018; Abeler, Nosenzo and Raymond, 2019). This feature has implications for many settings plagued by bureaucratic misconduct, where effective accountability mechanisms are either lacking or difficult to enforce. As far as I know, such reputational discipline effect is rarely explored at the local official level.

Third, this paper relates to the literature estimating the real impacts of corruption. The distortions created by such misconduct have long been recognized (Krueger, 1974; Shleifer and Vishny, 1993), but well-identified empirical evidence is still inadequate due to the difficulty in measuring corruption (Banerjee, Mullainathan and Hanna, 2012). As such, the literature mostly estimates the real impacts of relevant interventions targeting corruption. Extant research in this vein finds that curbing corruption improves resource allocation and spurs economic activity (Giannetti et al., 2021; Colonnelli and Prem, 2022). My paper studies a less examined aspect of bureaucratic misconduct—the manipulation of economic statistics. By diverting local officials' effort from economic development, such misconduct may be similarly costly. Curbing such misconduct thus improved local officials' performance and spurred economic activities. In some domains such as credit allocation and firm entry, the impacts are comparable to those of anti-corruption campaigns as found both in China and Brazil. Hence, manipulation of economic statistics should be given similar emphasis as corruption, although in reality its salience is ignored.

Finally, while this paper focuses on the bureaucracy, the manipulation of information by local officials echoes earnings manipulation by corporate managers (Stein, 1989; Fischer and Verrecchia, 2000; Goldman and Slezak, 2006; Benmelech, Kandel and Veronesi, 2010; Agarwal, Daniel and Naik, 2011; Ma, Pan and Stubben, 2020). In this vein, the findings have implications beyond the bureaucratic setting. Different from corporate managers, local officials often wield broader influence over the local economy (Xu, 2011). Thus, their manipulation carries widespread welfare consequences.

The remainder of the paper is organized as follows. Section 2 introduces the institutional background of GDP growth manipulation, the reform, and a conceptual framework illustrating the impacts of the reform. Section 3 describes the data and discusses the empirical strategy. Section 4 presents the main results on GDP growth manipulation, robustness tests, and mechanisms. Section 5 investigates the impact on local officials' effort in developing the economy. Section 6 concludes.

2 Background and conceptual framework

2.1 Institutional background

In China, GDP is calculated using the value-added method; that is, by summing the value added in all sectors in a region, with the county as the lowest level of regions for GDP calculation. The calculation is done by the local statistical bureaus, which are controlled by local officials in terms of personnel and funding. It is worth noting that a Chinese county is co-led by two leaders: the party secretary controlling personnel and other political affairs, and the magistrate running the economy (Xu, 2011). However, both leaders are evaluated heavily on GDP growth and hence have incentives to manipulate (Yao and Zhang, 2015). These leaders have a couple of ways to manipulate GDP: directly asking local statistical bureaus to make up numbers, requiring firms to overstate income or pay additional "tax" and return later, or double counting firms' non-local subsidiaries, among others.

In terms of the statistical reform, it was initiated by the National Bureau of Statistics (NBS), joint with other central authorities, in May 2009 with the goal of disciplining misreporting of local governments in processing statistical data. It mainly targeted local officials who falsified statistical data by themselves, forced or instructed other agents to manipulate, retaliated against those detecting manipulation, or failed to find severe distortion in local statistical data. The last clause means that local officials were still punished even if there was no evidence of their direct manipulation, alleviating concerns about the local statistical bureau acting as scapegoats upon detection. In addition, other agents participating in manipulation, such as the staff in local statistical bureaus, were also punished. The punishment was enforced by the relevant authorities (local officials superiors and the supervisory organs) and mostly took the form of reputational warning or demerit. Specifically, if a local official was found to manipulate economic statistics, a warning or demerit would be issued within the bureaucracy. In serious circumstances, demotion or dismissal would be issued.[1] Nevertheless, no legal actions were specified, which differentiates this reform from the commonly used random audits in Western countries that carry legal consequences (Avis, Ferraz and Finan, 2018).

In terms of the detection, both the local statistical bureau in each county and centrally managed survey teams in some counties, which I will describe their deployment shortly, were responsible. However, the local statistical bureau shared aligned incentives with local officials because they were appointed and funded by local officials; in contrast, the survey teams had a higher probability of detecting manipulation, as they were appointed and funded centrally. As emphasized in the literature, this type of independence is the key to the effectiveness of monitoring (Kofman and Lawarrée, 1993; Olken, 2007; Vannutelli, 2022). Upon detection, the survey teams could directly report to the NBS, and the NBS would take actions together with other authorities. The main takeaway so far is that the survey teams, which only existed in part of the counties, create the key source of variation in each county's exposure to the reform, enabling me to identify the reform' effect through difference-in-differences identification strategy.

What is crucial to my difference-in-differences strategy is when and how the survey teams were deployed. They were deployed by the NBS in 2005 in 40% of counties. Their initial job was to conduct sampling surveys to collect information on CPI, household income, grain output, and micro-firm dynamics. As these variables were frequently published and updated by the NBS, the survey teams could lighten the workload of the generally understaffed local statistical bureaus. Starting from 2009 when the aforementioned reform was

[1] Given that demotion or dismissal of local officials was extremely rare in reality, these punishments were unlikely to be enforced, which is confirmed by my subsequent analysis in which I show the reform had no effect on personnel turnovers and official characteristics.

launched, these teams also began to detect the manipulation of statistical data. While the NBS did not officially reveal the criteria regarding the selection of counties with these survey teams, the initial goal—to generate nationally representatively information—suggests that the allocation of these survey teams to counties should be largely orthogonal to local economic data manipulation. Furthermore, by checking local gazetteers, I find that most of these survey teams launched in 2005 were restructured from earlier rural survey teams set up in 1984. At that time, these earlier rural survey teams were mainly used to collect information on agricultural output, and counties with them were selected randomly within a province. This unique feature could allow me to further address potential concerns about endogenous allocation, by utilizing exogenous variations from the random assignment of these earlier rural survey teams. I leave the detailed discussion of these earlier rural survey teams in Section 4.2.3, where they are used as an instrument. Unless explicitly noted, the survey teams refer to those launched in 2005 in my subsequent analysis.

2.2 Conceptual framework

In this part, I will leverage a simple economic tournament model to illustrate the sources of GDP growth manipulation under China's unique promotion rule and generate some testable predictions on the effect of the reform. Similar to the game setting in Lazear and Rosen (1981), I consider a single-period tournament without discounting. There are two county leaders indexed by $i = 1, 2$ competing for promotion, which is decided by the principal (the upper-level government) based on their reported economic performance. Leader i can manipulate GDP growth with effort m_i and stimulate the economy with effort e_i, subject to a constraint $m_i + e_i \leq \bar{C}$.[2] Conceptually, the effort exerted in manipulating GDP growth involves time, energy, and resources spent on cooking the book or persuading potential dissenters, among others. The payoffs to these two types of effort are $h(\cdot)$ and $g(\cdot)$, respectively, which are increasing and concave. $h(\cdot)$ is concave because, at higher levels of manipulation, an additional unit increase in manipulation is more likely to be detected and to incur greater dissent, thus requiring more effort. Furthermore, I assume that to make the manipulation less detectable, a leader conducts manipulation simultaneously with stimulating the economy, instead of after observing true GDP growth at the end of the period.[3] Hence, the reported GDP growth is given by

$$G_i = h(m_i) + g(e_i) + \varepsilon_i, \quad \varepsilon_i - \varepsilon_{-i} \sim \mathrm{U}[-\frac{1}{2\phi}, \frac{1}{2\phi}]$$

where the reported GDP growth is the sum of the payoffs of both types of effort, plus an idiosyncratic shock ε_i. I assume the difference of the shocks between the two counties is

[2] While the manipulation of GDP growth may also be done by the staff in local statistical bureaus, I only model the behaviors of local officials as local statistical bureaus are controlled by local officials and thus act in concert with local officials.

[3] This is also supported by anecdote evidence showing that local officials asked firms to overstate income in the middle of a year.

uniformly distributed with mean 0 and density ϕ. Such distribution is known to all, but the realized values of the shocks are only known at the end of the period.

In addition, manipulating GDP would be detected with probability p, where $p \in (0,1)$ denotes the exogenous rate of identifying manipulation. As discussed in the previous section, a warning or demerit will be issued within the bureaucracy upon detection. For severe manipulation, a demotion or dismissal will be issued. Thus, the leader suffers from a direct reputational loss due to the exposure of manipulation within the bureaucracy, which takes the linear form of $\lambda h(m_i)$. This type of reputational cost in shaping truth-telling has both theoretical and empirical foundations (Benabou and Tirole, 2011; Gneezy, Kajackaite and Sobel, 2018; Abeler, Nosenzo and Raymond, 2019). In addition, the leader could also suffer from a promotion cost, which I will model in the promotion part. Finally, legal cost is not modelled as the reform involves no legal actions. Then leader i's payoff is given by:

$$U_i = \mathbb{1}_{\{i \text{ promoted}\}} u(R) + (1 - \mathbb{1}_{\{i \text{ promoted}\}}) u(r) + u(\Omega) - \mathbb{1}_{\{i \text{ detected}\}} \lambda h(m_i)$$

where $\mathbb{1}_{\{i \text{ detected}\}}$ is an indicator equal to 1 if leader i is detected for manipulation and 0 otherwise, and $\mathbb{1}_{\{i \text{ promoted}\}}$ is an indicator equal to 1 if leader i is promoted and 0 otherwise. The utility function $u(\cdot)$ is increasing and concave. Leader i receives reward R if promoted and r if not, where $R \gg r > 0$.[4]. To generate sharp predictions on GDP growth manipulation, I abstract from corruption by assuming that leader i extracts a fixed amount of rents Ω from the current office.[5]

Promotion rule The promotion rule posits that the principal promotes the county leader with the highest reported GDP growth. This rule has a widely-acknowledged theoretical foundation, as the high comparability across subnational units in China makes economic tournaments particularly suitable for promoting regional leaders (Maskin, Qian and Xu, 2000). It is also verified by a growing literature showing that GDP growth is positively related to local officials' career advancement at various levels of governments (Li and Zhou, 2005; Xu, 2011; Jia, Kudamatsu and Seim, 2015; Landry, Lü and Duan, 2018). In addition, scholarship on political selection in China also emphasizes the role of connections with upper-level leaders (Shih, Adolph and Liu, 2012; Jia, Kudamatsu and Seim, 2015; Meyer, Shih and Lee, 2016). I abstract from this factor as it mainly matters for promotion at higher level. In particular, Landry, Lü and Duan (2018) show that at the county level, GDP growth has a significant impact on promotion while political connections do not. To capture the potential punishment on promotion upon detection of manipulation, I assume that the principal subtracts an amount of $\delta h(m_i)$ from a leader's reported GDP growth. Here, δ represents a promotional punishment, but how large it is remains an empirical question. So

[4]The positive reward r captures the fact that in China most local officials would still serve in a similar position even if not promoted, instead of exiting the bureaucracy.

[5]I support such assumption by empirically showing that the reform did not affect corruption.

leader i is promoted if

$$G_i - \mathbb{1}_{\{i \text{ detected}\}} \delta h(m_i) > G_{-i} - \mathbb{1}_{\{-i \text{ detected}\}} \delta h(m_{-i})$$

Timing The timing of events in this tournament is summarized as follows:

1. Both leaders simultaneously choose effort in manipulating GDP growth and stimulating the economy, before knowing the realization of ε_i.
2. ε_i is realized and all uncertainty is resolved.
3. The principal detects manipulation, punishes the involved, and makes promotion decision based on the aforementioned promotion rule.

Equilibrium The equilibrium concept is a pure strategy Nash equilibrium.[6] To solve it, note that leader i maximizes expected payoff taking leader $-i$'s choice as given. As shown in Appendix A, through usual maximization (assume interior solution), one can solve for equilibrium m^* (effort in manipulating GDP growth) and e^* (effort in stimulating the economy):

$$\begin{aligned} m^* &= K\left[\frac{V\phi}{V\phi(1-p\delta)-\lambda p}\right] \\ e^* &= \bar{C} - m^* \end{aligned}$$

where $K(\cdot)$ is the inverse function of $h'(\cdot)/g'(\bar{C}-\cdot)$. V is equal to $u(R)-u(r)$, which measures the utility gains from promotion.

The reform's effect Conceptually, the reform could decrease local leaders' effort in GDP growth manipulation m^* in three ways. First, by exposing manipulation upon detection, the reform could increase local leaders' reputational cost of manipulation, which is captured by λp. Second, by imposing possible penalties on local officials' promotion prospects upon detection, the reform could increase the promotional cost of manipulation, which is captured by δp. Third, the reform may also decrease m^* by removing those involved in manipulation upon detection, leading to a change of the shape of $K(\cdot)$. Regarding local leaders' effort in developing the economy e^*, the reform could increase e^* as it would be relatively less costly to develop the economy.

To empirically test the reform's effect, recall that the reform mainly relied on the survey teams deployed in some counties to detect manipulation. In counties without the survey teams, the detection of manipulation relied on local statistical bureaus. As local statistical bureaus share aligned incentives with local officials and are also controlled by local officials in terms of personnel and funding, they are essentially dysfunctional in terms of detection.

[6] I focus on pure strategies as it is empirically obscure to interpret mixed strategies in manipulating GDP growth and stimulating the economy in a static game.

As can be seen from the expression of m^*, these counties would essentially be unaffected by the reform without effective detection (i.e., $p = 0$). One could thus test the reform's effect through a difference-in-differences framework by comparing counties with and without the survey teams before and after 2009.

3 Data and empirical strategy

3.1 Main data

Below, I briefly describe the main data used in this paper. More details about these data, along with descriptions of additional data, will be provided when the data first appears in the paper.

The list of counties with survey teams The list of counties with survey teams deployed in 2005 is collected from the annual reports published by various levels of statistical bureaus, and supplemented by local gazetteers.

County-level outcomes and covariates County-level data on GDP and other variables are collected from county statistical yearbooks. County-level data on harmonized nighttime light intensity are collected from Li et al. (2020). These data will be used to construct proxies for GDP growth manipulation. Other county data on demographic, economic, and geographic characteristics, which are used to conduct balance tests and serve as controls, are collected from multiple sources including the 2010 population census, the National Oceanic and Atmospheric Administration (NOAA), and the United States Geological Survey (USGS).

Local leader résumés Local leader résumés for the party secretary and magistrates are collected from various government websites, Baidu Baike (China's equivalent of Wikipedia), and occasionally complemented by online news reports. They are used to construct variables on leader characteristics.

County-level government annual work reports County-level government annual work reports, which outline a county's development policies, are collected from the government websites of each county. This data will be used to examine local officials' policy changes.

Bank credit County-level data on bank loans and branches for various types of banks is collected from the China Banking Regulatory Commission. This data will be used to examine credit allocation.

Firm entry Firm registration data, initially at the firm level for various types of firms (private firms, state-owned enterprises (SOEs), foreign-owned firms, and collectively owned firms), is collected from (Dong et al., 2021). This data will be used to examine firm entry.

Household surveys Household survey data (the China Family Panel Studies) is collected from the Institute of Social Science Survey (ISSS) maintained by Peking University. This survey is a nationally representative survey and will be used to examine citizens' attitudes towards local officials.

My main analysis focuses on a county-level panel from 2005 to 2018. To construct the county sample, I exclude the following special types of counties following the convention in the literature (Li, Lu and Wang, 2016; Chen et al., 2020): (1) counties in the four centrally-managed cities (*Beijing*, *Shanghai*, *Tianjin*, and *Chongqing*). These counties have a higher political status than others and, therefore, are not comparable to other counties; (2) urban districts, which are more developed economically but less independent administratively;[7] (3) counties in Tibet where data is unavailable; (4) counties outside mainland China.[8] I then define treatment counties as those with the surveys teams deployed in 2005 and control counties as those without. In the end, I have 1,779 counties in total, of which 40% are treated. The spatial distribution of treatment and control counties can be found in Figure 1, which is quite even across space.

3.2 Empirical strategy

Deriving estimation equation To derive the estimation equation, I incorporate GDP growth manipulation and the reform into the framework of Henderson, Storeygard and Weil (2012), who establish the positive relationship between nighttime light intensity and real economic activities. First, denote the reported GDP growth (with manipulation), true GDP growth (unobservable), and nighttime light intensity growth in county c and year t as z_{ct}, y_{ct}, and l_{ct}, respectively. Assume the degree of manipulation is m_{ct}. Then the GDP growth observed by the local statistical bureau is $z_{ct} - m_{ct}$ (without manipulation). According to Henderson, Storeygard and Weil (2012), the mapping from $z_{ct} - m_{ct}$ to y_{ct}, and the mapping from l_{ct} to y_{ct} can be written respectively as:

$$z_{ct} - m_{ct} = y_{ct} + \epsilon^{z}_{ct} \tag{1}$$

$$l_{ct} = \gamma y_{ct} + \epsilon^{l}_{ct} \tag{2}$$

Combining equations (1) and (2), the degree of manipulation can be written as:

$$m_{ct} = z_{ct} - \frac{1}{\gamma} l_{ct} + \epsilon^{m}_{ct} \tag{3}$$

where ϵ^{m}_{ct} is a combination of the error terms ϵ^{z}_{ct} and ϵ^{l}_{ct}. Then the difference-in-differences

[7] Both urban districts and counties are county-level divisions under a prefecture-level city. Some key differences are: (1) urban districts are the core areas of a prefecture-level city and are thus more developed; (2) urban districts have strong dependency on the city in terms of administrative functions (such as land development, urban planning, fiscal expenditure, etc.).

[8] Specifically, these include counties in Hongkong, Macau, and Taiwan. They are excluded due to institutional and administrative differences from mainland China.

equation to test the effects of the reform on GDP growth manipulation can be written as:

$$m_{ct} = \beta Treat_c \times Post_t + \delta_c + \lambda_t + \varepsilon^m_{ct} \tag{4}$$

where $Treat_c$ and $Post_t$ are dummy variables for treatment counties (the 40% aforementioned counties with survey teams deployed in 2005) and post-reform years (years after 2009), respectively. δ_c denotes county fixed effects, controlling for time-invariant factors at the county level that may correlate with the treatment or the outcome; λ_t denotes year fixed effects, controlling for time-varying shocks common to all counties. As one cannot directly observe m_{ct}, substituting equation (3) into equation (4) and rearranging generates:

$$\underbrace{z_{ct}}_{\text{Reported GDP Growth}} = \frac{1}{\gamma} \underbrace{l_{ct}}_{\text{Light Growth}} + \beta Treat_c \times Post_t + \delta_c + \lambda_t + \varepsilon^z_{ct} \tag{5}$$

where ε^z_{ct} is a combination of ϵ^m_{ct} and ε^m_{ct}. One can then estimate this equation. Since the treatment varies at the county level, I cluster the standard errors by county (Abadie et al., 2023) and assess robustness using alternative inference procedures, such as clustering at different levels (city and province), correcting for spatial correlation (Conley, 1999), and employing randomization inference (Young, 2019). I expect the coefficient of interest β to be negative, which implies that counties with the survey teams would engage in less manipulation relative to other counties after the reform. Note that both local statistical bureaus and the survey teams could engage in detecting manipulation after the reform, which may affect the interpretation of β. Under the assumption that local statistical bureaus are dysfunctional in detection, which is plausible as they are controlled by local officials, the coefficient β could be well interpreted as the overall effect of the reform. If this assumption is not true, the coefficient β is a lower bound of the effect of the reform, but this is still meaningful. In addition, β may also capture the lower bound effect in the presence of spillover effects among counties, which I will rule out empirically.

Identification concerns The identification assumption is that, reported GDP growth, after adjusting for light growth, should evolve in parallel between treatment and control counties in the absence of the reform in 2009. This assumption is essentially unverifiable. Pre-reform parallel trends between treatment and control counties, which is commonly estimated using event study specifications, can lend support to this assumption but cannot fully verify it. One still needs to address two types of concerns: first, the relationship between economic growth and light growth may differ across counties or years, which is specific to my setting; second, treatment counties may differ significantly from control counties ex ante, a common concern in difference-in-differences designs. To address the first concern, in the spirit of Martinez (2022), I will allow the effect of light to vary by a host of spatial and temporal characteristics to check the sensitivity of the estimates. For the second concern, while perfect covariate balance ex ante is not necessarily required in such designs, significant imbalance may cast doubt on the validity of using the control groups as counterfactual. To check this, Table 1 provides a balance test along various baseline county covariates, which shows that treatment counties were quite similar to control counties ex ante, except for the levels of population and GDP. The pre-deployment nature of the survey teams implies that such imbalance should not be endogenously related to the reform and therefore should be largely orthogonal to my outcomes. Indeed, there was no significant difference for pre-reform GDP growth and light growth, implying that treatment and control counties were similar in terms of GDP growth manipulation (or lack thereof). In the robustness checks, I also adopt various methods to address concerns about covariate imbalance, such as flexibly controlling for size and other baseline covariates, allowing for county-specific trends, and achieving covariate balance through entropy balancing (Hainmueller, 2012) and coarsened exact matching (Iacus, King and Porro, 2012). I further leverage institutional knowledge to design two additional tests to address concerns about unobserved heterogeneity: the first is a placebo event study around 2005 when the survey teams were launched but had not yet undertaken any disciplining actions; the second is an instrumental variable estimation using the randomly assigned rural survey teams in the 1980s as an instrument, which I will elaborate on later.

4 Results on GDP growth manipulation

4.1 Main results

Event study Figure 2 shows the dynamic effect of the reform estimated using an event study specification (6), with baseline county covariates included gradually from panel (a) to panel (d). The year before the reform, 2008, is omitted as the reference year. The coefficient estimates in the pre-reform period, namely, β_js for $j < 2009$, are essentially small in magnitude and statistically insignificant. F-tests of joint significance of all the pre-reform estimates generate p-values larger than 0.9 in all specifications, implying that the parallel trends assumption is plausibly satisfied. Note that recent econometric literature shows that this type of pretests may be underpowered to detect a diverging pre-trend (Roth, 2022). However, as I show in subsequent analysis, the results are robust to accounting for potential pre-trends using a couple of methods including an instrumental variable approach. In the post-reform period, there is an immediate and persistent negative effect, suggesting that the reform decreased GDP growth manipulation, which confirms the prediction from the conceptual framework. In Appendix Figure A1, I further show a decomposition of the effect of the reform, by checking the dynamic effect on reported GDP growth and light growth separately.[9] The results further confirm that the reform decreased GDP growth manipula-

[9]Specifically, I estimate the following equation, where Y_{ct} denotes either reported GDP growth or light growth:

$$Y_{ct} = \sum_{j=2005,\ j\neq 2008}^{j=2018} \beta_j Treat_c \times 1_{\{t=j\}} + \delta_c + \lambda_t + \varepsilon_{ct}$$

tion: there is a sharp drop in reported GDP growth but little change in light growth in the post-reform period.

$$ReportedGDPGrowth_{ct} = \alpha LightGrowth_{ct} + \sum_{j=2005,\ j\neq 2008}^{j=2018} \beta_j Treat_c \times 1_{\{t=j\}} + \delta_c + \lambda_t + \varepsilon_{ct} \quad (6)$$

Average effect Table 2 summarizes the dynamic treatment effect above into an average treatment effect. Column (1) reports the results using equation (5), controlling for only county and year fixed effects, and light growth. The point estimate on Treat × Post is negative and statistically significant (coef.=−0.751, s.e.=0.316), implying that relative to the control counties, treatment counties experienced a 0.751 percentage points drop in GDP growth manipulation after the reform. The estimate on light growth is also consistent with that in the literature (Martinez, 2022).[10] Through columns (2)-(4), I gradually introduce a set of baseline demographic, economic, and geographic controls (interacted with the post-reform dummy), which are presented in the balance tests. The precision of the estimates improves, although the size drops slightly. In the most stringent specification in column (4) with all the county controls, the estimate shows a 0.576 (s.e.=0.161) percentage points drop in GDP growth manipulation in treatment counties relative to control counties after the reform. This drop is also economically substantial, which amounts to 5.3% of the mean of the reported GDP growth. In sum, these findings suggest the effectiveness of the combination of monitoring and punishment in reducing bureaucratic misconduct in data processing, which resonates with the key insights from Becker and Stigler (1974). In Appendix Table A1, I also show the estimates on reported GDP growth and light growth separately, and find a similar decrease in reported GDP growth (coef.=−0.576, s.e.=0.161). The effect on light growth is small and statistically insignificant (coef.=0.254, s.e.=0.294). This means that the reform reduced GDP growth manipulation primarily through a reduction in reported GDP growth.[11]

Spillover Having established the negative effect of the reform on GDP growth manipulation, I turn to check if there exists any spillover effect, which could bias my baseline estimation even if the treatment is exogenous.[12] To this end, I estimate equation (7), where $Spillover_c$ denotes the strength of spillover to county c from other counties. Thus, β^{Direct} captures the direct effect of the reform while $\beta^{Spillover}$ captures the spillover effect. Following Avis, Ferraz and Finan (2018) and Huber (2023), I use the number of treatment counties among a county's neighbors to proxy for the strength of spillover to that county, where

[10]Specifically, in a similar specification in Martinez (2022)'s cross-country analysis, the coefficient estimate on light growth is about 0.027-0.039 (s.e.=0.006-0.007), and in my setting, it is about 0.017-0.023 (s.e.=0.005).

[11]If any, the positive yet insignificant effect on light growth implies my estimates may be a lower bound of the true impact on GDP growth manipulation.

[12]To the extent that there is a spillover effect to control counties, my estimates would underestimate the true reform effect.

neighbors are defined as other counties sharing a common boundary segment with that county.[13] The results are reported in Table 3. Column (1) reproduces the baseline estiamtes. In column (2), the estimated spillover effect is small and statistically insignificant (coef.=−0.01, s.e.=0.071). Considering the average number of treatment neighbors for a county is 2, such estimates imply that moving from a county with no treatment neighbors to the average county would decrease GDP growth manipulation by an additional 0.02 percentage points. Given the direct effect of about 0.58 percentage points, the spillover effect is thus economically negligible. In contrast, the direct effect remains virtually unchanged compared to the baseline effect in column (1). In the remaining two columns, I use dummies to indicate counties with at least one treatment neighbor or with treatment neighbors higher than the median, and the results have no substantial changes. In Appendix Table A2, I further show that the results are robust to using alternative definitions of neighbors or weighting the treatment neighbors by their sizes. The lack of a spillover effect can be well reconciled with two facts: (1) the coverage of treatment counties is fixed over time, as they hinges on the pre-deployed survey teams; (2) the evaluation of local officials may be among similar counties (Xu, 2011), namely, within either treatment or control counties.

$$\begin{aligned} ReportedGDPGrowth_{ct} = \alpha LightGrowth_{ct} + \ & \beta^{Direct} Treat_c \times Post_t \\ & + \beta^{Spillover} Spillover_c \times Post_t + \delta_c + \lambda_t + \varepsilon_{ct} \end{aligned} \tag{7}$$

4.2 Robustness checks

4.2.1 Alternative specifications

In this section, I show the robustness of the baseline results to a host of alternative specifications that alleviate the aforementioned identification concerns. The first concern is about heterogeneous mapping between nighttime light intensity and true economic activities. As previously discussed, the baseline equation (5) assumes a uniform and linear relationship between these two variables. However, such relationship may be nonlinear and change across counties or over years. To alleviate this concern, I allow the effect of light to: (1) be non-linear by including a 3rd-order polynomial of light; (2) vary by county longitude and latitude; (3) vary by county area; (4) vary by baseline GDP (5) vary by baseline population (6) vary by baseline urbanization rate; (7) vary by baseline economic structure (proxied by share of population in the primary and secondary sectors); (8) vary by year; (9) vary by province; (10) vary by both province and year; (11) vary by treatment status; (12) vary by both treatment status and linearly by year. To better examine the sensitivity to a specific modification, I estimate a variant of the baseline equation (5) each time according to one of the modifications above, and plot the results in panel (a) of Figure 3. The results are essentially unaffected by these alternative specifications, which suggests that my baseline

[13] To alleviate concerns that this number captures the geographic centrality of a county, I include neighbor number fixed effects interacted with the post-reform dummy throughout the estimation of spillover effect. Results are essentially unaffected by such fixed effects.

findings are not an artifact of heterogeneous light effect. In Panel (a) of Appendix Figure A2, I futher control for baseline county covariates and also estimate a specification incorporating all the modifications above. The patterns are similar.

The second concern is about covariate imbalance. As shown in Table 1, treatment counties were larger in size than control counties ex ante and therefore may differ significantly from control counties later on, leading to potential violations of the parallel trends assumption. To alleviate this concern, I estimate the following alternative specifications: (1) I flexibly control for size effects by including county size decile bin fixed effects interacted with year fixed effects, where county size is proxied by baseline GDP, population, or area; (2) I add county-specific time trends that allow treatment and control counties to be on differential linear trajectories (Angrist and Pischke, 2014). This could relax the identification assumption, although the precision of the estimates may decrease;[14] (3) I add province $\times$ year fixed effects. In this way, I am only comparing counties in the same province and year, and the covariates should be more balanced; (4) I include all the baseline county covariates interacted with year fixed effects to allow treatment and control counties to trend differentially depending on the covariates; (5) I select the most relevant covariates using the Double LASSO method (Belloni, Chernozhukov and Hansen, 2014); (6) I re-weight observations to make treatment and control counties similar in terms of observables using the entropy balancing method (Hainmueller, 2012); (7) I adopt the coarsened exact matching (CEM) method to match treatment counties to control counties within groups defined by all intersections of the deciles of baseline GDP, population, and area (Iacus, King and Porro, 2012). To better examine the sensitivity to a specific modification, I estimate a variant of the baseline equation (5) each time according to one of the modifications above, and plot the results in panel (b) of Figure 3.[15] These estimates have no substantial changes compared to the baseline estimate, implying that my findings are not driven by possible differential trends caused by covariate imbalance. I further control for baseline county covariates when appropriate in panel (b) of Appendix Figure A2 and find similar patterns.

I also address some other concerns in the remaining panels of Appendix Figure A2. Panel (c) shows that the results are not driven by a particular region, by conducting estimations leaving out each province individually. Panel (d) shows that the results are not driven by a few marginal counties by weighting the regression by county size (e.g., population or GDP). Panel (e) shows that the results are robust to alternative levels of clustering (by city or province) and spatial correlation correction (Conley, 1999).[16] Panel (f) shows that the results are robust to randomization inference, which has better finite sample properties and is also insensitive to high-leverage observations (Young, 2019).[17]

[14]Using linear time trends in DiD specifications could absorb part of the effect and the treatment variation, which leads to less precise estimates (Goodman-Bacon, 2021).

[15]Note that unlike in panel (a), here a specification incorporating all the modifications is unfeasible.

[16]For province-level clustering with a small number of clusters of 26, I also report the wild bootstrap p-value with 2,000 replications (Roodman et al., 2019). For Conley standard errors, I account for serial correlation spanning all years and spatial correlation within distances of 250 km, 500 km, 750 km, and 1,000 km.

[17]Following the recommendation by Young (2019), I use 2,000 permutations as the marginal gain from ad-

4.2.2 Placebo reform: The launch of the survey teams in 2005

I corroborate the previous results by conducting a placebo event study around 2005 when the survey teams were launched but had not conducted any disciplining actions. This could further help to examine if there existed any pre-trends. In addition, it could alleviate further concerns that the survey teams per se may affect the outcomes, even without the reform in 2009. The specification is equation (8), which is similar to the baseline event study specification in equation (6) except that the sample period here is from 2001 to 2008, with the year 2004 omitted as the reference year. Figure 4 shows the event study estimates. In contrast to the sharp drop in reported GDP growth after 2009 in Figure 2, there was no discernible change in reported GDP growth around 2005. This pattern persisted until 2009. Furthermore, there were no diverging trends before 2005. Such results suggest that the reform effect is unlikely to be driven by baseline differences between treatment and control counties or differential effects (net of the reform effect) generated by the survey teams per se.

$$
\begin{aligned}
ReportedGDPGrowth_{ct} = \alpha LightGrowth_{ct} + \sum_{j=2001,\ j\neq 2004}^{j=2008} \beta_j Treat_c \times 1_{\{t=j\}} \\
+ \delta_c + \lambda_t + \varepsilon_{ct} \qquad (8)
\end{aligned}
$$

4.2.3 Instrumental variable strategy

While the results presented so far could alleviate most concerns about covariate unbalance, unobserved heterogeneity is still possible. For example, to ensure the quality of information collected, counties with the survey teams may be those good counties with a potentially downward and time-varying trend in GDP growth manipulation. In this case, the OLS estimates would overestimate the true impact of the reform. This may be legitimate concern given the positive selection nature of most policy experimentations in China (Wang and Yang, 2021). To alleviate this concern, in this section I leverage a unique institutional feature to construct an instrumental variable for the treatment counties and conduct an instrumented difference-in-differences estimation.

Background In 1984, to gauge agricultural production, the National Bureau of Statistics (NBS) set up a group of teams called *rural survey teams* in part of the counties. At that time, China was essentially an agricultural country. The counties with these rural survey teams were chosen randomly within a province. In particular, the NBS adopted a commonly used probability sampling method called *systematic random sampling*. Under this sampling method, one first selects a random starting point in a sequence of counties and then chooses counties at fixed and periodic intervals. I collect the list of counties eventually selected from provincial gazetteers.[18] In theory, counties within the same province should have

ditional permutations is minimal. The randomization inference p-value from this exercise is 0.001.

[18] Gazetteers are called *Difangzhi* in Chinese and are a series of encyclopedias covering a wide range of

the same probability of being selected, leading to perfect within-province randomness of assignment of counties with the rural survey teams. In practice, the randomness may be affected by particular patterns in the county sequence or the limited number of counties in some provinces,[19] and I will provide several tests to check the randomness. In terms of specific work, these rural survey teams were guided by the NBS, but in terms of personnel and funding, they were controlled by local officials. Given the dramatic change in economic structures caused by market reform in recent years, in 2005, these rural survey teams were abolished, and most of them were restructured into more comprehensive and independent survey teams led solely by the NBS, which are the survey teams examined in the previous parts. In sum, the unique random assignment feature of these earlier rural survey teams suggests that they could be used as an instrument for treatment counties, with the validity formally examined below.

Relevance I define an instrumental variable $Treat_c^{1984}$, which is a dummy equal to 1 if county c had a rural survey team in 1984. Given the previous discussion, I expect this instrument to be strongly correlated with $Treat_c$. Panel (a) of Figure 5 shows the distribution of the rural survey teams launched in 1984. The significant overlap with the treatment counties suggests the high relevance of the instrument. To formally assess the strength of this instrument, I report the first-stage regression results in panel (a) of Table 4 according to the following equation:

$$\begin{aligned} Treat_c \times Post_t = \theta LightGrowth_{ct} + \ \tau Treat_c^{1984} \times Post_t \\ + \gamma_p \times Post_t + \delta_c + \lambda_t + \epsilon_{ct} \end{aligned} \tag{9}$$

I include province fixed effects γ_p interacted with the post-reform dummy $Post_t$ to account for the fact that the random assignment is stratified by province (Duflo, Glennerster and Kremer, 2007; Bruhn and McKenzie, 2009). The coefficient estimates on $Treat_c^{1984} \times Post_t$ are positive and highly statistically significant across specifications, indicating the instrument's strong relevance. Following the suggestion by Andrews, Stock and Sun (2019), I report the effective F-statistic to assess the strength of the instrument (Montiel Olea and Pflueger, 2013).[20] Across all specifications, the effective F-statistics are around 2,000, which far exceeds both the rule-of-thumb value of 10 and the 5% critical value of 37.4.(Montiel Olea and Pflueger, 2013). In addition, Lee et al. (2022) argue that inference relying on the first-stage F-statistic exceeding a certain threshold may still be distorted, unless the F-statistic is larger than 104.7.[21] In this sense, the large F-statistics in my case are sufficient to guarantee correct inference in the second stage. In addition to the high relevance, it is worth noting that

topics: history, geography, economics, politics, culture, social sciences, etc. They are compiled by local officials and noted literati in each county and updated every dozens of years.

[19] In my sample, the number of counties in each province ranges from 13 to 128, with an average of 80.

[20] In just-identified cases, the effective F-statistic is equivalent to the conventional Kleibergen-Paap F-statistic.

[21] With F-statistics smaller than 104.7, Lee et al. (2022) shows that an adjustment factor should be applied to inflate the second-stage standard errors to deliver correct inference, which is not needed in my case.

the coefficients remain highly stable when various county characteristics are included. This stability points to the random assignment nature of the instrument, which I will further test below.

Exogeneity For the instrument $Treat_c^{1984}$ to be valid, it must also be uncorrelated with any other determinants of my outcomes, except the treatment, namely, the launch of a survey team in 2005. While this condition is essentially untestable, the random assignment nature of the instrument suggests that this condition should be plausibly satisfied. To check this, I run separate univariate regressions of various baseline demographic, economic, and geographic county characteristics on the instrument with province fixed effects included. Panel (b) of Figure 5 plots the standardized coefficients from this exercise. These coefficients are not only statistically insignificant but also centered around zero with small magnitudes, strongly supporting the randomness of the instrument.

To provide further evidence, I compute the standardized differences between counties with and without the rural survey teams.[22] Since the randomization is stratified by province, I first calculate the standardized differences within each province and then calculate a weighted average using the number of counties within each province as weights. Appendix Figure A3 plots the distribution of the standardized differences across county characteristics. The absolute values of the standardized differences never exceed 8%, which is far below the threshold of 25% for covariate balance as suggested by Imbens and Rubin (2015) .

Despite the strong evidence on the randomness of the instrument, one remaining concern is about potential *legacy effect* generated by the rural survey teams. Specifically, these rural survey teams, albeit abolished in 2005, may still have a lasting impact on the outcomes examined in my study period. A possible cause is that local officials may develop a cautious character due to the existence of the rural survey teams. If this character is passed on to subsequent leaders and leads to less manipulation, then the exclusion restriction could be violated. Conceptually, I view this as very unlikely due to two facts. First, the rural survey teams did not conduct any disciplining actions and were essentially led by local officials. Second, it is very rare that two consecutive leaders could co-work in the same county for a long period. In most turnover years, the overlap of their reigns is no longer than a few months, meaning that the legacy effect, if any, is unlikely to be persistent.

Another possible cause of the legacy effect is that upper-level governments may acquire better information about the agricultural sector in the counties with the rural survey teams. If so, it would be difficult for these counties to manipulate data on agricultural production even after the abolition of the rural survey teams. However, this should lead to differential shares of the agricultural sectors, which is not the case as one could see from the balance test. This is likely due to the frequent turnovers of upper-level leaders and significant changes in economic structures, rendering information acquired decades ago less useful.

[22] The standardized difference between two groups within a province is calculated as the difference between the sample means normalized by the square root of the average of the sample variances.

While there may still exist other sources of the legacy effect, it is worth noting that the legacy effect could be *differenced out* by my DID strategy around 2009, as long as it did not change over time, which seems plausible. In other words, only time-varying legacy effect that is correlated with my outcomes would lead to a violation of the exclusion restriction. One way to rule out a time-varying legacy effect is to check whether there exist any pre-trends before 2009 using the reduced-form event study specification as below:

$$\begin{aligned} ReportedGDPGrowth_{ct} = \alpha LightGrowth_{ct} + \sum_{j=2005,\ j\neq 2008}^{j=2018} \beta_j Treat_c^{1984} \times 1_{\{t=j\}} \\ + \gamma_p \times Post_t + \delta_c + \lambda_t + \varepsilon_{ct} \end{aligned} \tag{10}$$

where the only difference of this specification with the baseline event study specification is that the $Treat_c$ there is replaced by the instrument $Treat_c^{1984}$. β_js are the coefficients of interest and are expected to be different from zero in the pre-reform period if there exists a time-varying legacy effect. Panel (c) of Figure 5 plots the coefficient estimates from this exercise, which are small and statistically insignificant in the pre-reform period. This rules out the possibility of a time-varying legacy effect directly affecting manipulation in my study period.

2SLS estimates Given the relevance and exogeneity of the instrument, I perform 2SLS estimation with the following second-stage equation:

$$\begin{aligned} ReportedGDPGrowth_{ct} = \vartheta LightGrowth_{ct} + \ \beta \widehat{Treat_c \times Post_t} \\ + \gamma_p \times Post_t + \delta_c + \lambda_t + \varepsilon_{ct} \end{aligned} \tag{11}$$

where all variables are defined as previously, and the coefficient of interest is β, which captures the average causal effect of the reform. To check the dynamics, I first conduct an instrumented event study and plot the coefficient estimates in panel (d) of Figure 5. Similar to the baseline OLS event study (Figure 2), the figure shows no pre-trends in the pre-reform period and a persistent drop in GDP growth manipulation in the post-reform period, albeit with less precision. To summarize the dynamic effect, I report the second-stage results estimated based on equation (11) in panel B of Table 4. The results are also similar to OLS estimates both in terms of economic magnitude and statistical significance. Considering the estimate in the last column with the inclusion of baseline county covariates interacted with the post-reform dummy, the estimate shows a 0.59 percentage points drop in GDP growth manipulation in treatment counties relative to control counties after the reform in 2009 (coef.=-0.590, s.e.=0.203), which accounts for 5.4% of the mean of reported GDP growth and is nearly identical to the baseline OLS magnitude of 5.3%. Such similar results imply that the baseline findings are unlikely to be confounded by selection bias. This is consistent with the aforementioned fact that treatment counties were determined several years before the reform; therefore, any preexisting differences between treatment and control counties should be largely orthogonal to GDP growth manipulation. In Appendix Table A1, I further show that the reduction in GDP growth manipulation is mainly driven by a reduction in reported GDP growth, by providing IV estimates on reported GDP growth and light growth separately.[23]

Difference between IV and OLS While the similar magnitudes found in the IV and OLS estimations imply that selection bias is a minimal concern, it could also be that the IV estimates only reflect the treatment effect on compliers, namely, counties affected by the instrument. In the presence of heterogeneous treatment effects across counties, the treatment effect on compliers may differ from that on the full sample.[24] I take several steps to alleviate this concern. First, I calcuate the proportion of compliers and profile their characteristics using the method developed by Marbach and Hangartner (2020). The detailed procedures are described in Appendix B. I find that about three quarters of the counties are compliers. These counties also look similar to the full sample on all observable dimensions as shown in Appendix Figure A4. Second, I adopt a reweighting method in Appendix Table A4 to adjust the OLS estimates to match the sample of compliers (Bhuller et al., 2020; Agan, Doleac and Harvey, 2023).[25] The reweighted OLS estimates are quite similar to the unweighted OLS estimates. These results suggest that the similarity between the IV and OLS estimates is unlikely to be confounded by heterogeneous treatment effects.

4.3 Mechanisms and alternative explanations

After demonstrating the robustness of the baseline findings, in this section I explore the underlying mechanisms. I show that the drop in GDP growth manipulation is most consistent with a reputational discipline effect generated by the reform on local officials. Other mechanisms outlined in the conceptual framework, such as promotional discipline effect and personnel changes, play a minimal role. Additionally, I rule out several alternative explanations that could generate similar patterns, namely, enhanced soft information, improved local statistical capacity, and other concurrent reforms that may also strengthen the monitoring of local officials.

[23]Specifically, the table shows a similar decrease in reported GDP growth (coef.=-0.592, s.e.=0.203). The effect on light growth, while positive, is small and statistically insignificant (coef.=0.162, s.e.=0.355).

[24]Under heterogeneity treatment effects, the monotonic condition is also required, which means that counties affected by the instrument should be affected in the same way. A testable prediction of this condition is that the first-stage results should hold qualitatively across all subgroups. In Appendix Table A3, I provide suggestive evidence that this is satisfied by showing that the first-stage results are similar across subgroups divided by the medians of baseline county covariates.

[25]Specifically, I split the sample into multiple groups of equal size based on the baseline county covariates, and then reweight the OLS estimates using the complier share in each group as weights. See Appendix Table A4 for details.

4.3.1 Reputational discipline effect as the key mechanism

As discussed in the conceptual framework in Section 2.2, the reform could decrease GDP growth manipulation through two broad mechanisms: a *discipline* effect and a *selection* effect. The discipline effect suggests that local officials would refrain from manipulation within their terms due to reputational and promotional costs. The selection effect implies a reduction in manipulation as a result of personnel changes among local officials, stemming from the dismissal of those found to be involved in manipulation or the appointment of more competent successors.

To distinguish between the discipline effect and the selection effect, I construct a trimmed sample in which I require local officials' terms to straddle 2009 in the treatment counties.[26] Given the lack of personnel changes in the treatment counties in this sample, one should observe a smaller reduction in GDP growth manipulation in this sample if the reform partly worked through a selection effect. As presented in column (2) of Appendix Table A5, the reduction manipulation (coef.=-0.617, s.e.=0.207) in this trimmed sample is similar to the reduction (coef.=-0.576, s.e.=0.161) in the full sample as shown in column (1). If anything, the reform effect is slightly larger in this trimmed sample. This means that the selection effect cannot explain the drop in manipulation. In column (3), I further require local officials to stay at least two years both before and after the reform in the treatment counties, and the estimate is similar. In Appendix Table A6, I also test the effects of the reform on personnel turnovers and personnel traits, which are generally small and statistically insignificant. In sum, these results imply a minimal role played by the selection effect. This null effect is consistent with the fact that only the most severe cases of manipulation would result in the dismissal of local officials, which is itself a rare occurrence.

Given the absence of a selection effect, the reform likely worked through a discipline effect, which could be further decomposed into reputational and promotional discipline effects. For the latter effect, one should expect the reduction in GDP growth manipulation to vary by local officials' promotion incentives. In contrast, the former effect should not lead to such heterogeneity. To distinguish between these two mechanisms, I start with testing whether the reform effect is larger for local officials with greater promotion incentives. I measure local officials' promotion incentives utilizing age restrictions on promotion. Specifically, county-level leaders are generally unable to be promoted to the next level once their ages reach 52 (Kou and Tsai, 2014), which creates a sharp drop in promotion incentives at this age cutoff. In addition, as a county is co-led by both the party secretary and the magistrate, I adopt a specification allowing for their incentives to separately affect GDP growth manipulation (Yao and Zhang, 2015).[27] As shown in Appendix Table A8, I do not find a sta-

[26] As a county is co-led by the party secretary and the magistrate, I require the terms of both types of officials to straddle 2009 when constructing the trimmed sample.

[27] Specifically, following Yao and Zhang (2015), I treat the party secretary and the magistrate in a county as if they work in two different but identical counties. Empirically, this means that for each county-year, I generate two parallel observations that are identical except for one distinction: one includes only the party secretary, while the other includes only the magistrate.

tistically significant differential effect for local officials older than 52 who would have lower promotion incentives. As another measure of promotion incentives, I also estimate local officials' ex ante likelihood of promotion based on their start ages, years of schooling, and political connections with upper-level leaders (Avis, Ferraz and Finan, 2018; Wang, Zhang and Zhou, 2020).[28] I still do not find a statistically significant differential effect for local officials with higher promotion incentives as shown in Appendix Table A8. These results mean that promotional punishment cannot explain the drop in GDP growth manipulation. This could be possible if in reality the survey teams were only able to detect manipulation but cannot precisely determine the total amount of manipulation. Consequently, imposing a promotion-based punishment becomes impractical.

Taken together, the evidence so far is most consistent with the reform generating a reputational discipline effect on local officials. This echoes the insights from a growing literature highlighting the role of reputational concerns in shaping truthtelling (Benabou and Tirole, 2011; Gneezy, Kajackaite and Sobel, 2018; Abeler, Nosenzo and Raymond, 2019). I further rule out some other explanations in the following sections.

4.3.2 Alternative explanations

Soft information The survey teams may assist the upper-level government in achieving soft information about the performance of local officials, thereby dampening the role of GDP growth in promotion (Hart, 1995; Aghion and Tirole, 1997; Stein, 2002). Consequently, local officials may be less inclined to manipulate GDP growth after the reform. To explore this possibility, I focus on counties closer to the upper-level government or counties where the leaders are socially connected to the upper-level government. As well documented in the literature, shorter distances or social connections could also facilitate the flow of soft information.[29] As a result, one should expect to see a smaller reform effect as the soft information provided by the survey teams should be less instrumental in such counties. I follow the literature to measure social connections using shared hometown or education background between county leaders and leaders in the upper-level government (Shih, Adolph and Liu, 2012; Jia, Kudamatsu and Seim, 2015; Fisman et al., 2020).[30] As shown in Appendix Table

[28]Specifically, I estimate the following Probit model based on local officials' ages when they started their terms, years of schooling, political connections with upper-level leaders, and all two-way interactions between these three variables. I estimate it separately for the party secretary and the magistrate.

$$\Phi^{-1}[P(Promotion)] = \beta_0 + \beta_1 StartAge_i + \beta_2 Education_i + \beta_3 Connection_i + \beta_4 StartAge_i \times StartAge_i \\ + \beta_5 Education_i \times Connection_i + \beta_6 StartAge_i \times Connection_i + \varepsilon_i$$

Note that other performance variables, such as GDP growth, are intentionally excluded from this regression, so the estimated probabilities capture the ex ante likelihood of promotion (Wang, Zhang and Zhou, 2020). The estimates are reported in Appendix Table A7. The estimates are similar if I instead use a linear probability model.

[29]See for example Agarwal and Hauswald (2010), Petersen and Rajan (2002), Bandiera, Barankay and Rasul (2009), and Fisman, Paravisini and Vig (2017).

[30]Specifically, I create a dummy variable named Connection that equals 1 if the party secretary or magistrate in a county shares the same hometown or educational background with upper-level leaders, and 0 otherwise.

A9, I do not find a statistically significant differential effect for these counties, suggesting that soft information is unlikely to be a driving force of the baseline results.

Statistical capacity Local statistical bureaus may improve their statistical capacity through interactions with the survey teams, and hence could more accurately measure economic activity (Martinez, 2022). This may also lead to a drop in the gap between reported GDP growth and light growth. Given the difficulty in directly measuring the statistical capacity of a county, I utilize an award from the National Bureau of Statistics (NBS) for counties' outstanding performance in coordinating and conducting economic censuses, which are initiated every 4 or 5 years by the NBS. The award involves no material rewards. Data on recipient counties of such award is collected from the NBS and is available for the years 2004, 2008, 2013, and 2018. In Appendix Table A10, I show that treatment counties did not receive more such award after the reform, suggesting that the baseline results are unlikely driven by improvement in local statistical capacity. This null effect is further supported by an event study presented in Appendix Figure A5.

Concurrent reforms The baseline results could be confounded by concurrent reforms that may also strengthen the monitoring of local officials. The inclusion of province $\times$ year fixed effects in the robustness checks could rule out all confounding reforms at the province level. In this section, I examine two prominent reforms at the county level. The first reform is the province-managing-county (PMC) reform since 2003 (Li, Lu and Wang, 2016). This PMC reform stipulated that the provincial government could bypass the prefectural government and directly administer the county government in fiscal matters (Fiscal PMC), or even in all aspects for a few counties (Full PMC). This data is collected from various government websites. The second reform pertains to the top-down inspections during the recent anti-corruption campaign launched in 2013. These inspections mainly focused on curbing corruption but may also create discipline effect on all aspects. This data is complied by Wang (2021), who collects the detailed timings and sites of inspections from government websites and newspapers. Appendix Figure A6 presents the rollout of these reforms. Appendix Table A11 shows that the results have no substantial changes after accounting for these reforms.

5 Results on effort reallocation

Up to now, I have shown that the reform could address the information problem faced by upper-level governments, namely, GDP growth manipulation by local governments. As outlined in the conceptual framework, an increase in the cost of manipulation should create a shift of local officials' effort from manipulation to economic development after the reform. In this section, I test this conjecture by empirically examining various outcomes reflecting local officials' development effort, which can be grouped into four domains: government policies, bank credit, firm entry, and citizen attitudes towards local governments.[31] Each of these four domains captures a certain aspect of local officials' effort and complements the others. To minimize the risk of finding false positives across multiple outcomes, I focus on relatively aggregate outcomes in each domain, and also correct for multiple hypothesis testing for the aggregate outcomes following Anderson (2008).[32] As I demonstrate below, the results across various outcomes provide consistent and complementary evidence on a beneficial shift in local officials' effort following the reform.

5.1 Government policies

My key measure of local officials' development effort is their emphasis on policies beneficial to the economy, considering the significant latitude they have in shaping local development (Xu, 2011). To construct the measure, I conduct a textual analysis of government annual work reports. These reports are typically issued by local governments at the beginning of each year and are subject to approval by the People's Congress at the same level through anonymous voting. Each report contains two parts: a summary of the government's achievements in the last year and a work plan for the year ahead, which contains detailed and well-structured development policies. The emphasis on each policy area could vary significantly both cross-sectionally and temporarily, as the reports are essentially at the discretion of local officials. Hence, the reports are well-suited to examine local government policy changes (Jiang, Meng and Zhang, 2019; Campante, Chor and Li, 2022).[33]

I create an original dataset on county-level government annual work reports collected from the official websites of each county. To ease both collection and computational burdens,[34] I randomly select three provinces: *Guangdong*, *Shaanxi*, and *Zhejiang*.[35] The final sample includes 97 counties from 2005 to 2018, corresponding to 1,155 reports and over 320,000 sentences in total. I then define four policies beneficial to the economy: *business attraction*, *infrastructure investment*, *market reform*, and *policy experimentation*, which are major

[31] One concern with these outcomes is that they could also be manipulated. This concern should be minimal for two reasons. First, local officials are not evaluated by these outcomes, so their incentives to manipulate them are minimal. Second, outcomes in the latter domains (bank credit, firm entry, and citizen attitudes) are largely immune to interference by local governments as they are collected and maintained by either the central authorities or independent entities.

[32] Due to the different time spans and identification strategies used, further aggregation of the outcomes in different domains is not conducted. Instead, I report the sharpened q-values for the aggregate outcomes in Appendix Table A22 to adjust for multiple hypothesis testing (Anderson, 2008). The findings are essentially unaffected by this correction.

[33] Jiang, Meng and Zhang (2019) use prefecture-level government work reports and an unsupervised Latent Dirichlet Allocation topic model to examine social welfare policies at the prefecture level. Campante, Chor and Li (2022) also utilize prefecture-level government work reports, employing both a dictionary and a supervised machine learning approach, to measure governments' emphasis on political stability.

[34] Unlike the provincial-level or prefecture-level government websites, the county government websites have no uniform layouts, and much of the collection has to be done manually, which is a laborious process.

[35] This is done using a simple random sampling method. Namely, each of the 26 provinces in my full sample is first assigned a unique number from 1 to 26, then three random numbers between 1 and 26 are generated using a random number generator without replacement. The provinces corresponding to these three numbers are the randomly selected provinces.

contributory factors to China's recent economic success and are also frequently mentioned in the reports (Li and Zhou, 2005; Xu, 2011; Jiang, Meng and Zhang, 2019).[36] In addition, to alleviate concerns about multiple hypothesis testing, I create a standardized index following Kling, Liebman and Katz (2007), which is constructed as follows. First, I standardize each of the four policies to have a mean of 0 and standard deviation of 1. Second, I take an equally weighted average of the four standardized measures. Third, I standardize the weighted average again to have a mean of 0 and standardized deviation of 1, which serves as my main outcome of interest in this section.

I adopt two approaches to measure the emphasis on each policy, as detailed in Appendix C. The first is a keywords frequency approach. To this end, I first define a list of keywords corresponding to each policy and then count the total number of mentions of these keywords in each report. The list of keywords for each policy can be found in Appendix Table A12. I then normalize the keyword count by the total number of words in each report to account for differential length of each report. My second approach is a supervised machine learning approach. To this end, I randomly select 25% of the sentences from all reports and manually label them as belonging to each policy or not. I then apply two commonly used machine learning algorithms: *random forest (RF)* and *support vector machine (SVM)* (Gentzkow, Kelly and Taddy, 2019). These algorithms predict a binary policy score at the sentence level. I then construct a report-level policy score by taking the average of these scores weighted by sentence length.

Table 5 reports the effect of the reform on local government policies. The estimates are similar across both the key words frequency policy measures (panel A) and the machine learning-based policy measures (RF in panel B and SVM in panel C). When evaluated using the standardized policy index in column (1), the estimates indicate a roughly 50% standard deviation increase in local governments' emphasis on policies conducive to economic development after the reform. The estimates on individual policies in columns (2)-(5) reveal that this policy shift is driven by increased emphasis on business attraction and market reform, instead of infrastructure investment and policy experimentation. The null effect on infrastructure investment is consistent with criticisms about the sustainability of infrastructure-driven growth, considering the already substantial infrastructure stock (Zilibotti, 2017) in the 2010s. The null effect on policy experimentation is consistent with the discouraging effect of improved economic statistics on policy experimentation (Binswanger and Oechslin, 2020).[37] The event study graphs in Figure 6 confirm these patterns, showing no pre-trends before the reform and a subsequent positive and sustained shift in both the policy index and the two individual policies (business attraction and market reform).

Overall, these results suggest a shift of government policies in directions conducive to

[36]For policy experimentation, Xu (2011) uses the launch of special economic zones (SEZs) in *Shenzhen* and *Zhuhai* as an example to illustrate the pivotal role played by policy experimentation in China's economic rise.

[37]Specifically, a local government undertaking a policy experiment is less likely to receive the *benefit of the doubt* if the true numbers revealed by improved economic statistics suggest a failure of past attempts (Binswanger and Oechslin, 2020).

economic development after the reform. One remaining concern is that this policy shift may only reflect local officials' visions instead of tangible actions. I will address this concern in subsequent sections by further demonstrating positive effects on bank credit, firm entry, and citizen attitudes towards local officials.

Robustness I conduct two robustness tests in the Appendix. First, I employ a randomization inference procedure with 2,000 permutations to alleviate concerns about the small sample size, which may lead to distortions in conventional inference (Young, 2019). As shown in Appendix Figure A8, the randomization inference p-values are similar to conventional p-values. Second, I present IV estimates in Appendix Table A13 (average effect) and Figure A9 (event study) using the aforementioned randomly assigned rural survey teams in 1984 as an instrument for the treatment. The IV estimates have no substantial changes compared to the OLS estimates.

5.2 Bank credit

In many parts of the world including China, one important way that local officials could affect the economy is through their intervention in financial markets, particularly in the allocation of bank credit to firms (La Porta, Lopez-de Silanes and Shleifer, 2002; Dinç, 2005; Carvalho, 2014; Ru, 2018; Cong et al., 2019). To corroborate the previous findings on improvements in government policies, I turn to examining bank credit allocation following the reform in this section.

I collect disaggregated data on bank loans and branches at the county level from the China Banking Regulatory Commission for various types of banks. One drawback of such data is that it only covers the period 2006 to 2011, which limits the study of long-term effect. Nevertheless, it would still be reassuring if one finds a short-term effect. I construct four measures of bank credit, which are *total amount of loans*, *amount of loans to small firms*, *number of firms granted loans*, and *number of bank branches with loan services*. My focus is on the total amount of loans, which can be viewed as an aggregate outcome on bank credit, thus alleviating issues with multiple hypothesis testing. As the distributions of these variables are highly skewed, I apply the inverse hyperbolic sine transformation to reduce influences from the tails of the skewed distributions, and check robustness using untransformed variables. This transformation approximates the logarithm transformation but is well defined at zero (Bellemare and Wichman, 2020).[38]

Panel A of Table 6 shows that the reform generated positive effects on all four measures of bank credit. While the estimates are less precise, the economic magnitudes are generally larger than 10% except for the number of bank branches with loan services. To benchmark such magnitudes: Colonnelli and Prem (2022) show that random audits on local governments in Brazil increased bank loans by about 3%. In addition, the effect on loans to small firms is the largest and statistically significant (coef.=0.263, s.e.=0.116). The role of small and

[38] The inverse hyperbolic sine transformation of a variable x is: $IHS(X) = \ln(x + \sqrt{x^2 + 1})$.

medium enterprises (SMEs) in economic development is well documented in the literature and has been instrumental in China's recent economic progress. However, it is also true that SMEs face severe credit constraints (Beck, Demirgüç-Kunt and Maksimovic, 2008; Ayyagari, Demirgüç-Kunt and Maksimovic, 2010). The significant increase in loans to small firms thus speaks to positive policy shifts by local governments. The event study graphs in Figure 7 further confirms such patterns and show no pre-trends before the reform.

One concern with the previous difference-in-differences estimates is that they may capture credit demand instead of a government-led credit supply. To address this concern, I further conduct a difference-in-difference-in-diffferences (DDD) estimation exploiting differential control of banks by local governments across counties. To this end, I construct a county-level index $GovernmentControl_c$ as follows:

$$GovernmentControl_c = \sum_b LoanShare^{pre}_{cb} \times GovernmentControl_b$$

where $LoanShare^{pre}_{cb}$ denotes the share of bank b in county c's loan market prior to the reform, and $GovernmentControl_b$ denotes bank b's degree of control by local governments. I set $GovernmentControl_b$ to be 1 for City Commercial Banks (CCBs) whose controlling shareholders are local governments, and 0 for other banks. The county-level index $GovernmentControl_c$ is further standardized to ease interpretation. I then estimate the following DDD specification:

$$\begin{aligned} CreditOutcome_{ct} = &\ \beta_1 Treat_c \times GovernmentControl_c \times Post_t \\ &+ \beta_2 Treat_c \times Post_t + \beta_3 GovernmentControl_c \times Post_t \\ &+ \delta_c + \lambda_t + \varepsilon_{ct} \end{aligned} \tag{12}$$

where β_1 is the coefficient of interest capturing the differential impact of the reform on credit outcomes across counties with varying preexisting government control over banks. Panel B of Table 6 reports the results. The estimates on $Treat_c \times GovernmentControl_c \times Post_t$ are large in magnitude and also generally statistically significant. In particular, the estimates suggest that moving from the average county to a county with one standard deviation higher government control over banks would increase the reform's effect on the total amount of loans by around one quarter (coef.=0.237, s.e.=0.103). The event study graphs in Figure 7 further confirm this pattern. In contrast, the estimates on $Treat_c \times Post_t$, while mostly positive, are small and statistically insignificant. These results are consistent with an expansion of credit supply from the government side.

Robustness I conduct two robustness tests in the Appendix. First, recent econometric literature shows that the coefficient estimates with log-like transformations of variables, including the inverse hyperbolic sine transformation, are sensitive to variable units (Mullahy and Norton, 2022; Chen and Roth, 2023).[39] In Appendix Table A15, I use untransformed variables. The findings still hold qualitatively. Second, I present IV estimates in Appendix Table A14 (average effect) and Figure A10 (event study) using the aforementioned randomly assigned rural survey teams in 1984 as an instrument for the treatment. The IV estimates are similar to the OLS estimates.

5.3 Firm entry

In China, establishing a firm involves lengthy bureaucratic procedures, over which local officials have substantial discretion (WorldBank, 2008; Jia, Lan and Padró i Miquel, 2021).[40] Therefore, increased development effort by local officials should manifest as higher firm entry. I test this conjecture in this section using the universe of firm registration data from Dong et al. (2021), which is available for the years 2005, 2010, and 2015.[41] For my main analysis, I aggregate the firm-level registration data at the county $\times$ year level and further exploit the microstructure of the data for robustness. I create four variables representing the number of registrations for each of the four firm types classified by ownership: *private firms*, *state-owned enterprises (SOEs)*, *foreign-owned firms*, and *collectively owned firms*. To reduce the number of tests, I also create an aggregate variable representing the total number of registrations for all firms, which serves as my primary focus. As before, I apply the inverse hyperbolic sine transformation to these five variables to reduce the influence of outliers in the tails of the skewed outcome distributions and check robustness using untransformed variables (Bellemare and Wichman, 2020).

Panel A of Table 7 reports the results. In aggregate, the reform significantly increased firm entry by about 5% (coef.=0.046, s.e.=0.026), as shown in column (1). To put this effect into perspective: Giannetti et al. (2021) show that China's 2013 anti-corruption campaign increased firm entry by 6.7% for a province-industry that was initially one standard deviation more corrupt than the average. Although the specification differs, it nonetheless provides reassurance that my estimate is of considerable economic significance. When examining the effect by ownership in the remaining columns, I also find increased entry for private firms (coef.=0.048, s.e.=0.027), SOEs (coef.=0.169, s.e.=0.062), and foreign firms (coef.=0.041, s.e.=0.051), consistent with their relatively higher productivity and pivotal role in economic growth (Song, Storesletten and Zilibotti, 2011).[42] In contrast, the effect on collectively

[39] This concern is particularly prominent when variables are frequently observed at zero (Mullahy and Norton, 2022). However, in my case, all four variables have positive values for more than 93% of observations.

[40] In the World Bank's Doing Business report, China was ranked 151st out of 182 countries in 2008 in terms of the ease of starting a business. This ranking takes into account factors such as the number of procedures, time spent on registration, and cost relative to income (WorldBank, 2008).

[41] The authors collected the data by web scraping an online system called the National Enterprise Credit Information Publicity System (NECIPS) (see https://www.gsxt.gov.cn/index.html). In China, a newly established firm needs to register at the local Administration for Industry and Commerce by providing detailed information such as firm name and address. After approval, the relevant information is publicized on the NECIPS.

[42] The positive and larger effect on SOEs may seem puzzling as they are generally considered less productive, but SOE productivity has been converging with that of private firms after nearly a decade of productivity-

owned firms is negative, consistent with the fact that these firms are inefficient, although the estimate is imprecise (coef.=−0.035, s.e.=0.062). The event study graphs in Figure 8 further confirm these patterns and show that the increase in firm entry happened immediately after the reform, although the limitation of the data prevents me from examining the pre-trends.

To further tighten identification and alleviate concerns about pre-trends, I exploit the microstructure of the data to conduct a standard regression discontinuity (RD) design across county borders. I use towns as the unit of observation and collapse the firm-level registration data at the town × year level.[43] Each town is then assigned to the nearest county border. In cases where a county shares its border with multiple counties, the county's border is divided into multiple segments so that there is only one county on each side of a border segment. I then focus on towns along county borders with different treatment statuses on each side. I estimate the following local linear regression with a uniform kernel:

$$FirmEntry_i = \beta_1 Treat_i + \beta_2 Treat_i \times Distance_i + \beta_3 Distance_i + \delta_{b(i)} + \varepsilon_i \quad (13)$$
$$s.t. \quad -h < Distance_i < h$$

where $FirmEntry_i$ denotes the number of firm registrations in town i, either aggregate or by ownership.[44] $Treat_i$ is equal to 1 if town i is located in a treatment county and 0 otherwise. $Distance_i$ is the distance from the centroid of town i to the nearest county border, and is negative if $Treat_i$=0. To ensure that treatment and control towns are comparable, I include county border fixed effects $\delta_{b(i)}$ to restrict the comparison to be within a narrowly defined geographic area. I estimate the equation using the optimal bandwidth h proposed by Calonico, Cattaneo and Titiunik (2014) with standard errors clustered at the county border level.

The identification assumption of this RD design is that all other factors affecting firm entry should evolve smoothly across county borders, except for the reform. To check this, Appendix Figure A12 conducts a balance test on preexisting town covariates, which show no significant jumps at county borders. In addition, RD designs utilizing administrative borders may suffer from compound treatments issues if there are other institutional or regulatory differences across the borders (Keele and Titiunik, 2015). However, as long as these potential differences are not related to firm entry, the RD design is still valid. To check this, Appendix Figure A12 further presents a placebo RD estimation using pre-reform firm entry data and finds no discontinuities at county borders. The RD estimates on post-reform firm entry are presented in Figure 8 and Panel B of Table 7. Consistent with the DID estimates,

enhancing reforms in the state sector since the late 1990s (Hsieh and Song, 2015). See Appendix Figure A7 for the dynamics of firm TFP by ownership.

[43] To deal with the relatively frequent consolidations of towns, I map all the outcomes and covariates to towns based on a 2010 map. On average, a county contains about 15 towns in 2010.

[44] Bellemare and Wichman (2020) suggest using the inverse hyperbolic sine transformation for large values of outcomes (e.g., larger than 10). Given that the number of firm entry at the town level is in general smaller than 10, I use the raw number here.

the RD estimates show that the reform boosted firm entry, especially for those with higher productivity, which further corroborates the previous findings of greater effort exerted by local officials in economic development after the reform.

Robustness I conduct some tests to check the robustness of the findings in this section. First, for the DID estimation, I show in Appendix Table A16 that the estimates are robust to using untransformed firm entry data. I also present IV estimates in Appendix Table A17 (average effect) and Appendix Figure A11 (event study) using the aforementioned randomly assigned rural survey teams in 1984 as an instrument for the treatment. The patterns are similar. Second, for the RD estimation, I show in Appendix Table A18 that the estimates are robust to alternative bandwidth (Imbens and Kalyanaraman, 2012), quadratic polynomial, and triangular kernel. I also show in Appendix Figure A13 that the RD estimates are robust to a randomization inference procedure to address concerns about inference in RD designs (Ganong and Jäger, 2018).[45]

5.4 Citizen attitudes

Finally, increased development effort by local officials should be reflected in the attitudes of citizens towards local governments. To test this conjecture, I use survey data from the China Family Panel Studies (CFPS) and pool three waves, 2012, 2014 and 2016, together.[46] I focus on citizens' *trust in local officials* and *evaluation of local government performance*, two commonly used measures of the performance of government officials (Bertrand et al., 2020; Martinez-Bravo et al., 2022).[47] For the former, the survey asks: "To what extent do you trust local officials?" The answer is an integer from 0 to 10 with larger values denoting higher trust. For the latter, the survey asks: "What is your overall evaluation of the county government's achievements last year?" The answer is one of the following: significant achievement, some achievement, not much achievement, no achievement, worse than before. To ease interpretation, I create a dummy variable indicating some or significant achievement.[48] To tighten identification, I create two placebo variables utilizing another two questions. The first question asks: "Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?" The answer is binary, either yes or no, and I use a dummy variable to denote yes. The second question describes a hypothetical person with minor health issues and then asks: "What do you think

[45]Specifically, I keep only county borders across which there are no treatment variations, and then randomly create placebo treatment variation to each border. Based on these placebo borders, I then re-estimate equation 13. This process is repeated for 2,000 times.

[46]The baseline wave in 2010 is not used as it may take time for both local governments to take actions and the citizens to change attitudes. In addition, the two key variables I examine below are missing in the 2010 wave.

[47] One issue with these variables is that citizens may not express their opinions faithfully. To alleviate this issue, I drop the top decile of citizens who show the highest concerns about the survey, which are observed and recorded by the investigators. Results are similar if I instead control for citizens' concerns.

[48]Because of the categorical nature of the answer to the second question, I do not summarize these two questions into one. I instead correct for multiple hypothesis testing in Appendix Table A22 and the findings still hold.

about the health condition of the person?" The answer is one of the following: extremely healthy, very healthy, relatively healthy, average, not healthy. As before, I create a dummy variable indicating extremely or very healthy.

Due to the lack of pre-reform survey data, I follow the empirical strategy in Duflo (2001) to estimate a cohort DiD specification that utilizes two sources of variation: (1) treatment counties versus control counties; (2) most affected cohorts versus less unaffected cohorts within the same county. The latter source of variation is built on insights in the psychology and political science literature that citizens' political attitudes are most permeable during teenage years and keep stable since one's 30s (Wolfinger and Rosenstone, 1980; Krosnick and Alwin, 1989). I create three cohort groups: those born in the 1970s, in the 1980s, and in the 1990s.[49] The 1970s cohort are the unaffected group as they would be older than 30 during the reform period. The 1990s cohort are the most affected group as they would be younger than 30 during the reform period, and thus, their political attitudes would be most permeable. The 1980s cohort are defined as the less affected group because some of them would be older than 30 during the reform period. The estimation equation is:

$$Y_{icg} = \sum_{k=1980s,1990s,k\neq 1970s} \beta_k Treat_c \times 1_{\{g=k\}} + \delta_c + \lambda_g + W_i\Omega + X_{cg}\Psi + \varepsilon_{icg} \tag{14}$$

where Y_{icg} denotes attitude measures for citizen i living in county c and born in cohort $g \in \{1970s, 1980s, 1990s\}$. The 1970s cohort is the omitted reference group. W_i denotes a set of citizen controls, including years of schooling and its square, age and its square, dummy for male, dummy for living in the urban area, and dummy for survey wave. δ_c and λ_g are county and cohort fixed effects, which help to partial out county- and cohort-specific time-invariant confounding factors, respectively. To the extent that the treatment $Treat_c$ may be correlated with county characteristics, which may have differential impacts on citizen attitudes, I include in X_{cg} the same set of county controls as before, interacted with cohort fixed effects.

Table 8 presents the results, which are also visualized in Figure 9. Panel A examines the effect on citizens' trust in local officials. As shown in column (1), the reform significantly increased the 1990s cohort's trust in local officials (coef.=0.452, s.e.=0.154). In terms of magnitude, the estimate indicates a 9.3% increase in trust relative to the mean trust across all three cohorts. In contrast, there is no change in citizens' trust in most people as shown in column (2). The estimates are not only small in magnitude but also statistically insignificant. When examining the effect by survey wave in columns (3)-(5), I find a similar patter. In addition, the estimates, albeit with less precision, indicate that the reform changed citizens' attitudes in only two years. This may seem striking considering that attitudes often change

[49] Older cohorts, such as those born in the 1950s and the 1960s, are not used as controls in my main analysis as they grew up in turbulent times when China suffered from several catastrophic events (e.g., the Great Famine and the Cultural Revolution). These events may affect trust formation (Chen and Yang, 2015; Bai and Wu, 2020), making them less comparable to younger cohorts. However, as shown in Appendix Table A19, results are similar if I include these older cohorts as controls.

gradually, but can be reconciled with the immediate drop in GDP growth manipulation after the reform as shown previously. Panel B examines the effect on citizens' evaluation of local government performance. As shown in column (1), the reform positively shifted the younger cohorts' attitudes, especially for the 1990s cohort (coef.=0.052, s.e.=0.020). Relative to the sample mean, the estimate indicate a 6.3% increase for the 1990s cohort. The placebo estimates on citizens' evaluation of others' health are again small in magnitude and statistically insignificant as shown in column (2). The remaining columns examine the effect by survey wave and show that the improvement was concentrated in the 2016 wave, which is reasonable as it takes time for local governments to make tangible achievements. Taken together, these results bolster the previous findings that local officials exerted more effort in developing the economy after the reform.

Robustness I provide several tests to check the robustness of these findings. First, I adopt an alternative definition of affected and unaffected cohorts in Appendix Table A19, where affected cohort consist of only those born in the 1990s and unaffected cohort consist of those born in the 1970s or older. Second, I control for citizens' media access in Appendix Table A20, which could alleviate the concern that treatment counties may be better at propaganda that could disproportionately affect the young cohort if they have greater access than other cohorts.[50] Third, I provide IV estimates using the aforementioned randomly assigned rural survey teams in 1984 as an instrument for the treatment, which are presented in Appendix Figure A14 and Appendix Table A21. The results are essentially unchanged.

5.5 Alternative explanations

The results across the four domains examined above are consistent with greater development effort exerted by local officials after the reform. In this section, I examine a few alternative explanations that may generate observationally equivalent results.

Fiscal transfers Local officials' differential performance may result from varying fiscal transfers from upper-level governments, given the well-documented development impacts of such transfers (Litschig and Morrison, 2013; Corbi, Papaioannou and Surico, 2019). While I cannot directly test this possibility due to the lack of data on county-level fiscal transfers after 2007,[51] I provide some suggestive evidence showing that this is unlikely. Conceptually, there are two plausible causes of varying fiscal transfers from upper-level governments after the reform. First, the upper-level government may decrease the transfers as an implicit punishment of subordinates' misconduct, but this would then work against from finding a

[50] Specifically, I include three variables constructed from the survey regarding media access: (1) the number of days political news was accessed via television in the last week; (2) the number of days political news was accessed via Internet in the last week; (3) whether you have posted comments related to political issues and major national events on Internet in the past 12 months.

[51] Data on county-level fiscal transfers could be collected from the China Prefecture, City, and County Public Finance Statistics published by the Ministry of Finance of China (Jia, Liang and Ma, 2021). However, the publication of such data stopped after 2007.

positive impact on local officials' performance. Second, the upper-level government may increase the transfers if better data provided by the survey teams increased the "creditworthiness" of their subordinate counties. If this is true, then one should also find a similar increase in the transfers around the launch of the survey teams in 2005, as they collected additional information about the county, which should play a similar credit-enhancing role. To check this, I collect data on fiscal transfers up to 2007 from the China Prefecture, City, and County Public Finance Statistics. As shown in Appendix Figure A15 and Appendix Table A23, I find little change in fiscal transfers after 2005.

Policy diffusion The arrival of the survey teams may facilitate policy diffusion, either from upper-level governments or among similar local governments (Shipan and Volden, 2008; Wang and Yang, 2021; DellaVigna and Kim, 2022), which may eventually improve local officials' performance. Conceptually, as the survey teams were deployed years before the reform, one should see a pre-trend in local officials' performance if the previous findings are driven by policy diffusion, which is not the case. To further rule out the possibility of varying policy diffusion after the reform, I examine the similarity of local government work reports across counties, with the premise that greater policy diffusion after the reform should lead to a convergence of these reports among treatment counties. Specifically, I estimate:

$$Similarity_{ijt} = \beta Treat_{ij} \times Post_t + \delta_{ij} + \lambda_{it} + \gamma_{jt} + \varepsilon_{ijt} \tag{15}$$

where ij indicates county pairs ($i \neq j$), with ij being equivalent to ji . $Similarity_{ijt}$ denotes the pairwise textual similarity of government work reports, which is calculated following Kelly et al. (2021) and described in Appendix C. $Treat_{ij}$ equals 1 if both i and j are treatment counties, and 0 otherwise. I include county $\times$ year fixed effects (λ_{it} and γ_{jt}) and county pair fixed effects (δ_{ij}) to account for county-specific traits and inherent differences between counties in a pair, respectively. Standard errors are two-way clustered by both counties in a pair. The results are reported in Appendix Table A24. Column (1) considers all county pairs to account for both types of policy diffusion. Column (2) considers county pairs within the same province to examine policy diffusion among local governments. Column (3) considers county pairs spanning different provinces to examine policy diffusion from the central government. The results are generally small and statistically insignificant.[52] Appendix Figure A16 provides the event study graphs, further confirming the absence of policy diffusion.

Short termism Local officials may shift effort to other short-termist behaviors after the reform, such as prioritizing growth over other factors (e.g., social welfare and environmental protection) or overleverage through shadow banking (Xiong, 2018). As a result, one may also observe improvements in local officials' performance in economic development after the reform. Appendix Table A25 and Appendix Figure A17 examine local officials' emphasis

[52] If anything, the results suggest a slight decrease in pairwise similarity of county work reports for county pairs across provinces (coef.=-0.002, s.e.=0.001), but is small in magnitude relative to the mean (0.45).

on social welfare and environmental protection in government work reports, using both a keywords frequency approach and a machine learning approach, as described in Appendix C. The estimates are generally small and statistically insignificant. I then examine local government debt. To this end, I collect data on bond issuance by local government financing vehicles (LGFVs) from the Wind database, which serves as a proxy for local government debt.[53] The earliest issuance at the county level was in 2009. As such, I collapse the bond issuance data by county and estimate a cross-sectional regression using the randomly assigned rural survey teams in 1984 as the instrument for treatment counties. Appendix Table A26 reports the estimates, which are small and statistically insignificant.

Corruption A reduction in corruption, which is possible if the reform increased the *perceived cost* of all types of misconduct, may also generate observationally equivalent results given the distorting effect of corruption on the economy (Giannetti et al., 2021; Colonnelli and Prem, 2022). I examine this possibility using a comprehensive dataset on corruption convictions complied by Wang and Dickson (2022).[54] The dataset contains 10,797 corruption convictions from 2005 to 2016, with a vast majority (10,788) happening after 2012 when China's anti-corruption campaigns began. The few convictions (9) before 2012 were likely caused by the lack of enforcement instead of less corruption, and are dropped from my analysis. I then collapse the data by county and estimate a cross-sectional regression using the randomly assigned rural survey teams in 1984 as the instrument for treatment counties. The results are reported in Appendix Table A27. Throughout the table, I include the number of anti-corruption inspections using the data from Wang (2021) to address concerns about differential anti-corruption enforcement (Glaeser and Saks, 2006; Zhu, 2017). The estimates are generally small and statistically insignificant, both in aggregate (column 1) and by corruption types (columns 2-3).[55] To alleviate concerns about potential lags between corruption and subsequent convictions, which may result in a null reform effect if all convictions reflecting corruption before 2009, I examine more recent convictions in columns (4)-(6), namely, those in 2015 and 2016.[56] The results have no substantial change. To further corroborate these findings, I also utilize the previous CFPS survey and equation (14) to estimate the effect of the reform on citizens' perceived corruption about the government.[57] I still find no effect as shown in Appendix Figure A18 and Appendix Table A28.

[53]In China, local governments are prohibited from borrowing from banks or issuing bonds directly (Huang, Pagano and Panizza, 2020). Instead, they could set up LGFVs and then issue bonds through them, usually with land offered by local governments as collateral. In addition, these LGFVs could also borrow from banks, but such loans are usually not disclosed.

[54]Wang and Dickson (2022) collect the data from Tencent—the largest Internet company in China. In 2011 Tencent launched a searchable online database of all corruption convictions across China, and the authors scraped the website in August 2016. Unfortunately, the website is closed currently.

[55]The major type of corruption contains bribery and appropriation of public property. The remaining types include other misbehaviours that are also considered as corruption in China, such as sex scandals (but no data manipulation).

[56]The corruption convictions in 2015 and 2016 account for 61% of all convictions.

[57]The survey question asks: "In general, how serious do you think the problem of government corruption is in our country?" The answer ranges from 0 to 10, with larger integers denoting higher perceived corruption. This question is only available in the 2014 and 2016 waves.

Leveraging China's 2009 reform punishing economic data manipulation and counties' quasi-random reform exposure, I provide rich causal evidence showing that the reform led to: (1) a decrease in GDP growth manipulation amounting to 5% of reported GDP growth, driven by a *reputational discipline* effect; (2) an increase in politicians' development effort manifested in both policy rhetoric and downstream impacts, consistent with an *effort reallocation* effect. These results highlight the relevance of reputational punishment in weak institutional settings and the cost of overlooked bureaucratic misconduct distinct from corruption.

6 Conclusion

Manipulation of official statistics by government agents is a common phenomenon in the world, as evident in both anecdote evidence and academic research. A direct consequence of such manipulation is information distortion within the bureaucracy. The indirect consequences, however, are not well recognized. This paper focuses on the China setting to provide causal evidence on how to reduce such manipulation and whether reducing it could generate economic benefits. Utilizing multiple datasets, a unique reform targeting economic data manipulation, and multiple identification strategies including an instrumented difference-in-differences design, this papers shows that a combination of top-down monitoring and punishment could effectively reduce such manipulation. Moreover, curbing such manipulation further elicited local officials' development effort. As such, the reform generated significant downstream impacts. In certain domains such as credit allocation and firm entry, the impacts are comparable to those of an anti-corruption campaign documented in the literature both in China and Brazil (Giannetti et al., 2021; Colonnelli and Prem, 2022).

These striking downstream impacts provide a new perspective to understand how individual local officials could affect the macro-level economic outcomes. Unlike political corruption, which could directly affect resource allocation and further stifle economic development (Krueger, 1974; Shleifer and Vishny, 1993), economic data manipulation has a more subtle impact—it hurts the economy by inducing an unfavorable shift of local officials' effort. This subtle impact may explain why such manipulation is not well recognized as a detrimental bureaucratic misconduct, despite its ubiquity. In this vein, the findings have implications for the design of relevant interventions targeting such manipulation.

This study has some limitations, and I outline some directions for future research. First, the question of how such manipulation affects the entire economy is essentially a general equilibrium question. The research design in this paper, by construction, may only estimate the lower bound and partial equilibrium effect. Future work may explore this impact using structural approaches. Second, future work could enrich our understanding of the impacts of such manipulation by studying the effect on firm performance using firm census data.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. "When should you adjust standard errors for clustering?" *The Quarterly Journal of Economics*, 138(1): 1–35.

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for truth-telling." *Econometrica*, 87(4): 1115–1153.

Agan, Amanda, Jennifer L Doleac, and Anna Harvey. 2023. "Misdemeanor prosecution." *The Quarterly Journal of Economics*, 138(3): 1453–1505.

Agarwal, Sumit, and Robert Hauswald. 2010. "Distance and private information in lending." *The Review of Financial Studies*, 23(7): 2757–2788.

Agarwal, Vikas, Naveen D Daniel, and Narayan Y Naik. 2011. "Do hedge funds manage their reported returns?" *The Review of Financial Studies*, 24(10): 3281–3320.

Aghion, Philippe, and Jean Tirole. 1997. "Formal and real authority in organizations." *Journal of political economy*, 105(1): 1–29.

Al-Ubaydli, Omar, John A List, and Dana Suskind. 2019. "The science of using science: Towards an understanding of the threats to scaling experiments." National Bureau of Economic Research.

Anderson, Michael L. 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American statistical Association*, 103(484): 1481–1495.

Andrews, Isaiah, James H Stock, and Liyang Sun. 2019. "Weak instruments in instrumental variables regression: Theory and practice." *Annual Review of Economics*, 11: 727–753.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2014. *Mastering'metrics: The path from cause to effect.* Princeton university press.

Avis, Eric, Claudio Ferraz, and Frederico Finan. 2018. "Do government audits reduce corruption? Estimating the impacts of exposing corrupt politicians." *Journal of Political Economy*, 126(5): 1912–1964.

Ayyagari, Meghana, Asli Demirgüç-Kunt, and Vojislav Maksimovic. 2010. "Formal versus informal finance: Evidence from China." *The Review of Financial Studies*, 23(8): 3048–3097.

Bai, Liang, and Lingwei Wu. 2020. "Political movement and trust formation: Evidence from the Cultural Revolution (1966–76)." *European Economic Review*, 122: 103331.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2009. "Social connections and incentives in the workplace: Evidence from personnel data." *Econometrica*, 77(4): 1047–1094.

Banerjee, Abhijit, Sendhil Mullainathan, and Rema Hanna. 2012. "Corruption." National Bureau of economic research.

Becker, Gary S, and George J Stigler. 1974. "Law enforcement, malfeasance, and compensation of enforcers." *The Journal of Legal Studies*, 3(1): 1–18.

Beck, Thorsten, Asli Demirgüç-Kunt, and Vojislav Maksimovic. 2008. "Financing patterns around the world: Are small firms different?" *Journal of financial economics*, 89(3): 467–487.

Bellemare, Marc F, and Casey J Wichman. 2020. "Elasticities and the inverse hyperbolic sine transformation." *Oxford Bulletin of Economics and Statistics*, 82(1): 50–61.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies*, 81(2): 608–650.

Benabou, Roland, and Jean Tirole. 2011. "Laws and norms." National Bureau of Economic Research.

Benmelech, Efraim, Eugene Kandel, and Pietro Veronesi. 2010. "Stock-based compensation and CEO (dis) incentives." *The Quarterly Journal of Economics*, 125(4): 1769–1820.

Bertrand, Marianne, Robin Burgess, Arunish Chawla, and Guo Xu. 2020. "The glittering prizes: Career incentives and bureaucrat performance." *The Review of Economic Studies*, 87(2): 626–655.

Besley, Timothy J, Robin Burgess, Adnan Khan, and Guo Xu. 2021. "Bureaucracy and development."

Bhuller, Manudeep, Gordon B Dahl, Katrine V Løken, and Magne Mogstad. 2020. "Incarceration, recidivism, and employment." *Journal of Political Economy*, 128(4): 1269–1324.

Binswanger, Johannes, and Manuel Oechslin. 2020. "Better statistics, better economic policies?" *European Economic Review*, 130.

Bossuroy, Thomas, Clara Delavallade, and Vincent Pons. 2019. "Biometric tracking, healthcare provision, and data quality: experimental evidence from tuberculosis control." National bureau of economic research.

Bruhn, Miriam, and David McKenzie. 2009. "In pursuit of balance: Randomization in practice in development field experiments." *American economic journal: applied economics*, 1(4): 200–232.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik. 2014. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica*, 82(6): 2295–2326.

Campante, Filipe R, Davin Chor, and Bingjing Li. 2022. "The Political Economy Consequences of China's Export Slowdown." Working Paper.

Carvalho, Daniel. 2014. "The real effects of government-owned banks: Evidence from an emerging market." *The Journal of Finance*, 69(2): 577–609.

Chen, Jiafeng, and Jonathan Roth. 2023. "Logs with zeros? Some problems and solutions." Working Paper.

Chen, Yi, Ziying Fan, Xiaomin Gu, and Li-An Zhou. 2020. "Arrival of young talent: The send-down movement and rural education in China." *American Economic Review*, 110(11): 3393–3430.

Chen, Yuyu, and David Yang. 2015. "Historical traumas and the roots of political distrust: Political inference from the Great Chinese Famine." *Available at SSRN 2652587*.

Colonnelli, Emanuele, and Mounu Prem. 2022. "Corruption and Firms." *The Review of Economic Studies*, 89(2): 695–732.

Cong, Lin William, Haoyu Gao, Jacopo Ponticelli, and Xiaoguang Yang. 2019. "Credit allocation under economic stimulus: Evidence from China." *The Review of Financial Studies*, 32(9): 3412–3460.

Conley, Timothy G. 1999. "GMM estimation with cross sectional dependence." *Journal of econometrics*, 92(1): 1–45.

Corbi, Raphael, Elias Papaioannou, and Paolo Surico. 2019. "Regional transfer multipliers." *The Review of Economic Studies*, 86(5): 1901–1934.

DellaVigna, Stefano, and Woojin Kim. 2022. "Policy diffusion and polarization across US states." National Bureau of Economic Research.

Dinç, I Serdar. 2005. "Politicians and banks: Political influences on government-owned banks in emerging markets." *Journal of financial economics*, 77(2): 453–479.

Dong, Lei, Xiaohui Yuan, Meng Li, Carlo Ratti, and Yu Liu. 2021. "A gridded establishment dataset as a proxy for economic activity in China." *Scientific Data*, 8(1): 1–9.

Duflo, Esther. 2001. "Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment." *American economic review*, 91(4): 795–813.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using randomization in development economics research: A toolkit." *Handbook of development economics*, 4: 3895–3962.

Fischer, Paul E, and Robert E Verrecchia. 2000. "Reporting bias." *The Accounting Review*, 75(2): 229–245.

Fisman, Raymond, Daniel Paravisini, and Vikrant Vig. 2017. "Cultural proximity and loan outcomes." *American Economic Review*, 107(2): 457–492.

Fisman, Raymond, Jing Shi, Yongxiang Wang, and Weixing Wu. 2020. "Social ties and the selection of China's political elite." *American Economic Review*, 110(6): 1752–1781.

Ganong, Peter, and Simon Jäger. 2018. "A permutation test for the regression kink design." *Journal of the American Statistical Association*, 113(522): 494–504.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as data." *Journal of Economic Literature*, 57(3): 535–574.

Ghanem, Dalia, and Junjie Zhang. 2014. "'Effortless Perfection:'Do Chinese cities manipulate air pollution data?" *Journal of Environmental Economics and Management*, 68(2): 203–225.

Giannetti, Mariassunta, Guanmin Liao, Jiaxing You, and Xiaoyun Yu. 2021. "The externalities of corruption: Evidence from entrepreneurial firms in China." *Review of Finance*, 25(3): 629–667.

Glaeser, Edward L, and Raven E Saks. 2006. "Corruption in america." *Journal of public Economics*, 90(6-7): 1053–1072.

Gneezy, Uri, Agne Kajackaite, and Joel Sobel. 2018. "Lying aversion and the size of the lie." *American Economic Review*, 108(2): 419–453.

Goldman, Eitan, and Steve L Slezak. 2006. "An equilibrium model of incentive contracts in the presence of information manipulation." *Journal of Financial Economics*, 80(3): 603–626.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277.

Greenstone, Michael, Guojun He, Ruixue Jia, and Tong Liu. 2022. "Can technology solve the principal-agent problem? Evidence from China's war on air pollution." *American Economic Review: Insights*, 4(1): 54–70.

Hainmueller, Jens. 2012. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies." *Political Analysis*, 20(1): 25–46.

Hart, Oliver. 1995. *Firms, contracts, and financial structure.* Clarendon press.

Henderson, J Vernon, Adam Storeygard, and David N Weil. 2012. "Measuring economic growth from outer space." *American Economic Review*, 102(2): 994–1028.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design." *The Journal of Law, Economics, and Organization*, 7(special_issue): 24–52.

Hsieh, Chang-Tai, and Zheng Michael Song. 2015. "Grasp the large, let go of the small: The transformation of the state sector in China." National Bureau of Economic Research.

Huang, Yi, Marco Pagano, and Ugo Panizza. 2020. "Local crowding-out in China." *The Journal of Finance*, 75(6): 2855–2898.

Huber, Kilian. 2023. "Estimating general equilibrium spillovers of large-scale shocks." *The Review of Financial Studies*, 36(4): 1548–1584.

Iacus, Stefano M, Gary King, and Giuseppe Porro. 2012. "Causal inference without balance checking: Coarsened exact matching." *Political analysis*, 20(1): 1–24.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal bandwidth choice for the regression discontinuity estimator." *The Review of Economic Studies*, 79(3): 933–959.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Jia, Junxue, Xuan Liang, and Guangrong Ma. 2021. "Political hierarchy and regional economic development: Evidence from a spatial discontinuity in China." *Journal of Public Economics*, 194: 104352.

Jiang, Junyan, Tianguang Meng, and Qing Zhang. 2019. "From Internet to social safety net: The policy consequences of online participation in China." *Governance*, 32(3): 531–546.

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

Jia, Ruixue, Xiaohuan Lan, and Gerard Padró i Miquel. 2021. "Doing business in China: Parental background and government intervention determine who owns busines." *Journal of Development Economics*, 151: 102670.

Keele, Luke J, and Rocio Titiunik. 2015. "Geographic boundaries as regression discontinuities." *Political Analysis*, 23(1): 127–155.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Measuring technological innovation over the long run." *American Economic Review: Insights*, 3(3): 303–320.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz. 2007. "Experimental analysis of neighborhood effects." *Econometrica*, 75(1): 83–119.

Kofanov, Dmitrii, Vladimir Kozlov, Alexander Libman, and Nikita Zakharov. 2023. "Encouraged to Cheat? Federal Incentives and Career Concerns at the Sub-national Level as Determinants of Under-Reporting of COVID-19 Mortality in Russia." *British Journal of Political Science*, 53(3): 835–860.

Kofman, Fred, and Jacques Lawarrée. 1993. "Collusion in hierarchical agency." *Econometrica: Journal of the Econometric Society*, 629–656.

Kou, Chien-wen, and Wen-Hsuan Tsai. 2014. ""Sprinting with small steps" towards promotion: solutions for the age dilemma in the CCP cadre appointment system." *The China Journal*, 71(1): 153–171.

Krosnick, Jon A, and Duane F Alwin. 1989. "Aging and susceptibility to attitude change." *Journal of Personality and Social Psychology*, 57(3): 416.

Krueger, Anne O. 1974. "The political economy of the rent-seeking society." *The American economic review*, 64(3): 291–303.

Landry, Pierre F, Xiaobo Lü, and Haiyan Duan. 2018. "Does performance matter? Evaluating political selection along the Chinese administrative ladder." *Comparative Political Studies*, 51(8): 1074–1105.

La Porta, Rafael, Florencio Lopez-de Silanes, and Andrei Shleifer. 2002. "Government ownership of banks." *The Journal of Finance*, 57(1): 265–301.

Lazear, Edward P, and Sherwin Rosen. 1981. "Rank-order tournaments as optimum labor contracts." *Journal of Political Economy*, 89(5): 841–864.

Lee, David S, Justin McCrary, Marcelo J Moreira, and Jack R Porter. 2022. "Valid t-ratio Inference for IV." *American Economic Review*.

Levinsohn, James, and Amil Petrin. 2003. "Estimating production functions using inputs to control for unobservables." *The review of economic studies*, 70(2): 317–341.

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

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

Litschig, Stephan, and Kevin M Morrison. 2013. "The impact of intergovernmental transfers on education outcomes and poverty reduction." *American Economic Journal: Applied Economics*, 5(4): 206–240.

Li, Xuecao, Yuyu Zhou, Min Zhao, and Xia Zhao. 2020. "A harmonized global nighttime light dataset 1992–2018." *Scientific Data*, 7(1): 1–9.

Lyu, Changjiang, Kemin Wang, Frank Zhang, and Xin Zhang. 2018. "GDP management to meet or beat growth targets." *Journal of Accounting and Economics*, 66(1): 318–338.

Mahadevan, Meera. 2023. "The price of power: Costs of political corruption in Indian electricity." *Working Paper*.

Ma, Matthew, Jing Pan, and Stephen R Stubben. 2020. "The effect of local tournament incentives on firms' performance, risk-taking decisions, and financial reporting decisions." *The Accounting Review*, 95(2): 283–309.

Marbach, Moritz, and Dominik Hangartner. 2020. "Profiling compliers and noncompliers for instrumental-variable analysis." *Political Analysis*, 28(3): 435–444.

Martinez-Bravo, Monica, Gerard Padró i Miquel, Nancy Qian, and Yang Yao. 2022. "The rise and fall of local elections in China." *American Economic Review*, 112(9): 2921–58.

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

Maskin, Eric, Yingyi Qian, and Chenggang Xu. 2000. "Incentives, information, and organizational form." *The Review of Economic Studies*, 67(2): 359–378.

Meyer, David, Victor C Shih, and Jonghyuk Lee. 2016. "Factions of different stripes: gauging the recruitment logics of factions in the reform period." *Journal of East Asian Studies*, 16(1): 43–60.

Montiel Olea, José Luis, and Carolin Pflueger. 2013. "A robust test for weak instruments." *Journal of Business & Economic Statistics*, 31(3): 358–369.

Mullahy, John, and Edward C Norton. 2022. "Why transform Y? A critical assessment of dependent-variable transformations in regression models for skewed and sometimes-zero outcomes." National Bureau of Economic Research.

Olken, Benjamin A. 2007. "Monitoring corruption: evidence from a field experiment in Indonesia." *Journal of Political Economy*, 115(2): 200–249.

Petersen, Mitchell A, and Raghuram G Rajan. 2002. "Does distance still matter? The information revolution in small business lending." *The journal of Finance*, 57(6): 2533–2570.

Roodman, David, Morten Ørregaard Nielsen, James G MacKinnon, and Matthew D Webb. 2019. "Fast and wild: Bootstrap inference in Stata using boottest." *The Stata Journal*, 19(1): 4–60.

Roth, Jonathan. 2022. "Pretest with caution: Event-study estimates after testing for parallel trends." *American Economic Review: Insights*, 4(3): 305–322.

Ru, Hong. 2018. "Government credit, a double-edged sword: Evidence from the China Development Bank." *The Journal of Finance*, 73(1): 275–316.

Shih, Victor, Christopher Adolph, and Mingxing Liu. 2012. "Getting ahead in the communist party: explaining the advancement of central committee members in China." *American political science review*, 106(1): 166–187.

Shipan, Charles R, and Craig Volden. 2008. "The mechanisms of policy diffusion." *American journal of political science*, 52(4): 840–857.

Shleifer, Andrei, and Robert W Vishny. 1993. "Corruption." *The quarterly journal of economics*, 108(3): 599–617.

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

Stein, Jeremy C. 1989. "Efficient capital markets, inefficient firms: A model of myopic corporate behavior." *The quarterly journal of economics*, 104(4): 655–669.

Stein, Jeremy C. 2002. "Information production and capital allocation: Decentralized versus hierarchical firms." *The journal of finance*, 57(5): 1891–1921.

Vannutelli, Silvia. 2022. "From Lapdogs to Watchdogs: Random Auditor Assignment and Municipal Fiscal Performance." National Bureau of Economic Research.

Wallace, Jeremy L. 2016. "Juking the stats? Authoritarian information problems in China." *British Journal of Political Science*, 46(1): 11–29.

Wang, Erik H. 2021. "Frightened mandarins: the adverse effects of fighting corruption on local bureaucracy." *Forthcoming, Comparative Political Studies*.

Wang, Shaoda, and David Y Yang. 2021. "Policy experimentation in China: The political economy of policy learning." National Bureau of Economic Research.

Wang, Yuhua, and Bruce J Dickson. 2022. "How corruption investigations undermine regime support: Evidence from China." *Political Science Research and Methods*, 10(1): 33–48.

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

Wolfinger, Raymond E, and Steven J Rosenstone. 1980. *Who votes?* Yale University Press.

WorldBank. 2008. *Doing business 2009.* The World Bank.

Xiong, Wei. 2018. "The mandarin model of growth." National Bureau of Economic Research.

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

Yao, Yang, and Muyang Zhang. 2015. "Subnational leaders and economic growth: evidence from Chinese cities." *Journal of Economic Growth*, 20(4): 405–436.

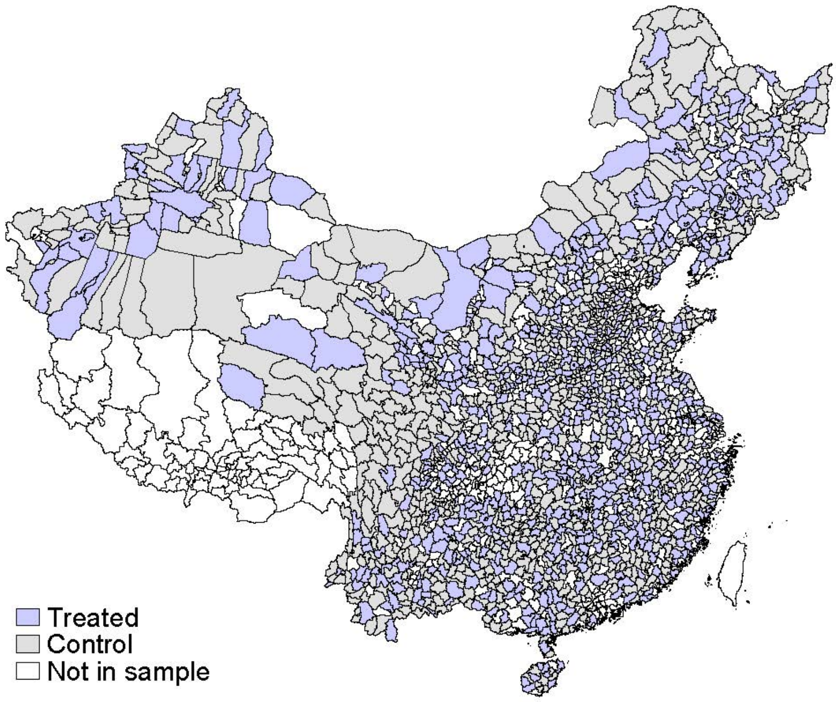
Young, Alwyn. 2019. "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results." *The quarterly journal of economics*, 134(2): 557–598.

Zhu, Boliang. 2017. "MNCs, rents, and corruption: Evidence from China." *American Journal of Political Science*, 61(1): 84–99.

Zilibotti, Fabrizio. 2017. "Growing and slowing down like China." *Journal of the European Economic Association*, 15(5): 943–988.

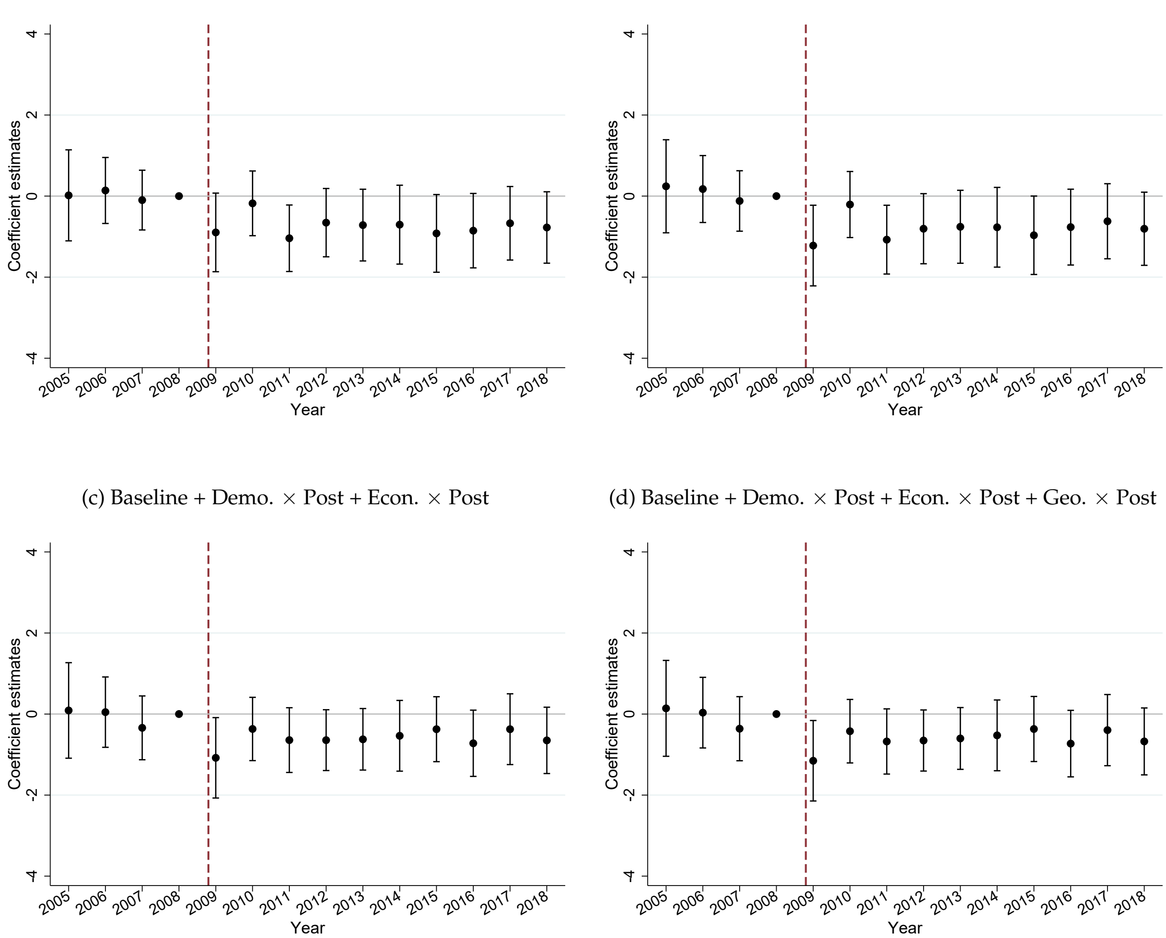
Figures and Tables

Figure 1: Distribution of treatment and control counties



Notes: This figure shows the distribution of the treatment and control counties, where treatment counties are those with the surveys teams deployed in 2005 and control counties are those without. Counties not in sample include: (1) counties in the four centrally-managed cities (*Beijing*, *Shanghai*, *Tianjin*, and *Chongqing*); (2) urban districts, which are more developed economically but less independent administratively; (3) counties in Tibet where data is unavailable; (4) counties outside mainland China.

Figure 2: Dynamic effect on GDP growth manipulation



(a) Baseline specification

(b) Baseline + Demo. × Post

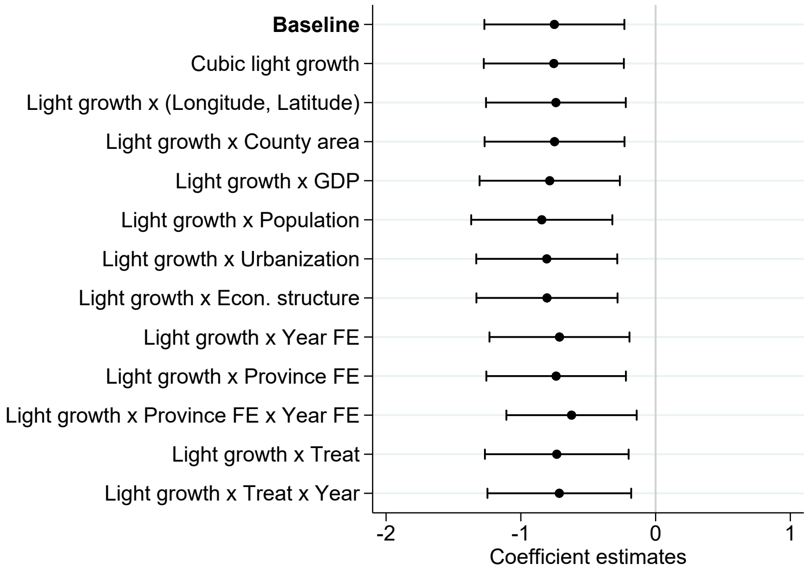
(c) Baseline + Demo. × Post + Econ. × Post

(d) Baseline + Demo. × Post + Econ. × Post + Geo. × Post

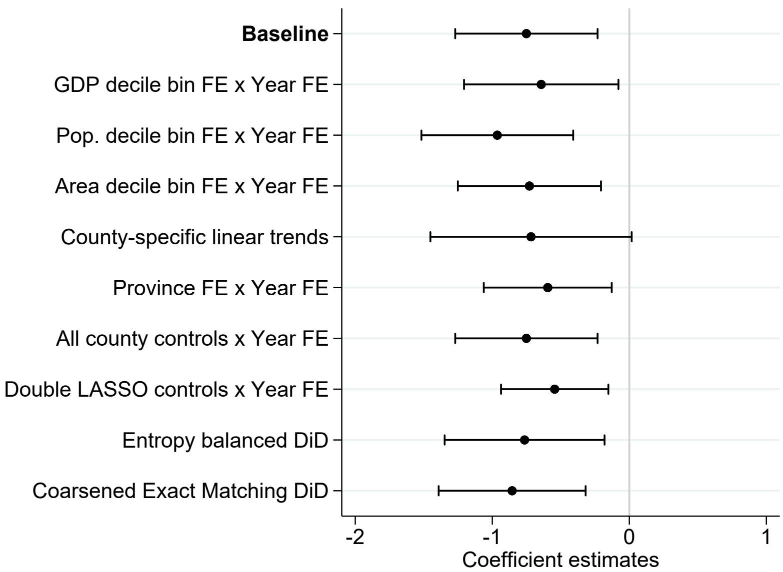
Notes: This figure shows the dynamic effect of the reform on GDP growth manipulation using the baseline event study specification (equation 6) with baseline county covariates included gradually. The year 2008, one year before the reform in 2009, is omitted as the reference year. Standard errors used to construct the 90% confidence intervals, which are denoted by the spikes, are clustered at the county level.

Figure 3: Sensitivity to alternative specifications

(a) Flexible effects of light growth



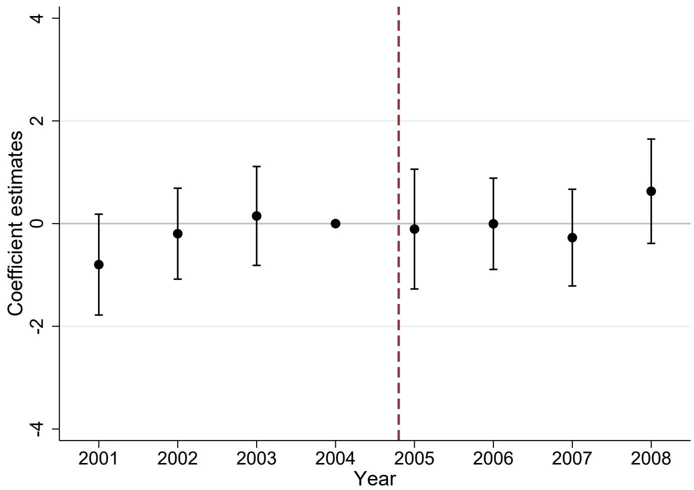
(b) Addressing covariate imbalance



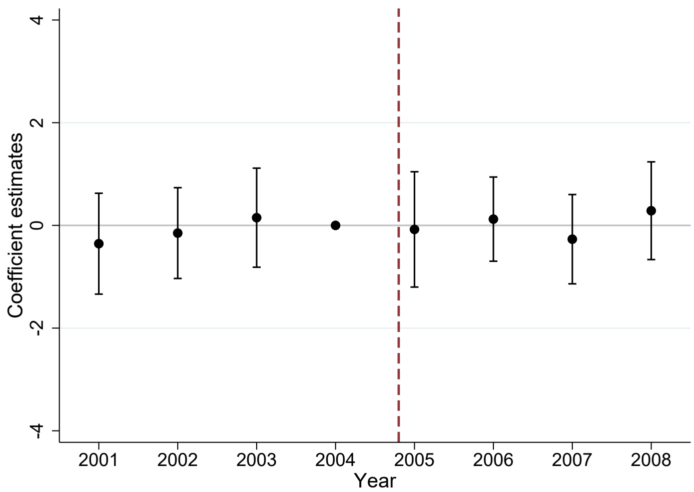
Notes: This figure checks the sensitivity of the baseline results to alternative specifications, and is created by estimating variants of the baseline equation (5). Specifically, panel (a) addresses the concern that the mapping between light growth and economic growth may not be uniform across counties or years, by allowing the mapping to vary flexibly; panel (b) addresses the concern that the results may be confounded by covariate imbalance between treatment counties and control counties, by directly controlling for the sources of imbalance or achieving covariate balance through entropy balancing and coarsened exact matching, among others. Standard errors used to construct the 90% confidence intervals, which are denoted by the spikes, are clustered at the county level.

Figure 4: Using the launch of the survey teams in 2005 as a placebo

(a) Baseline specification

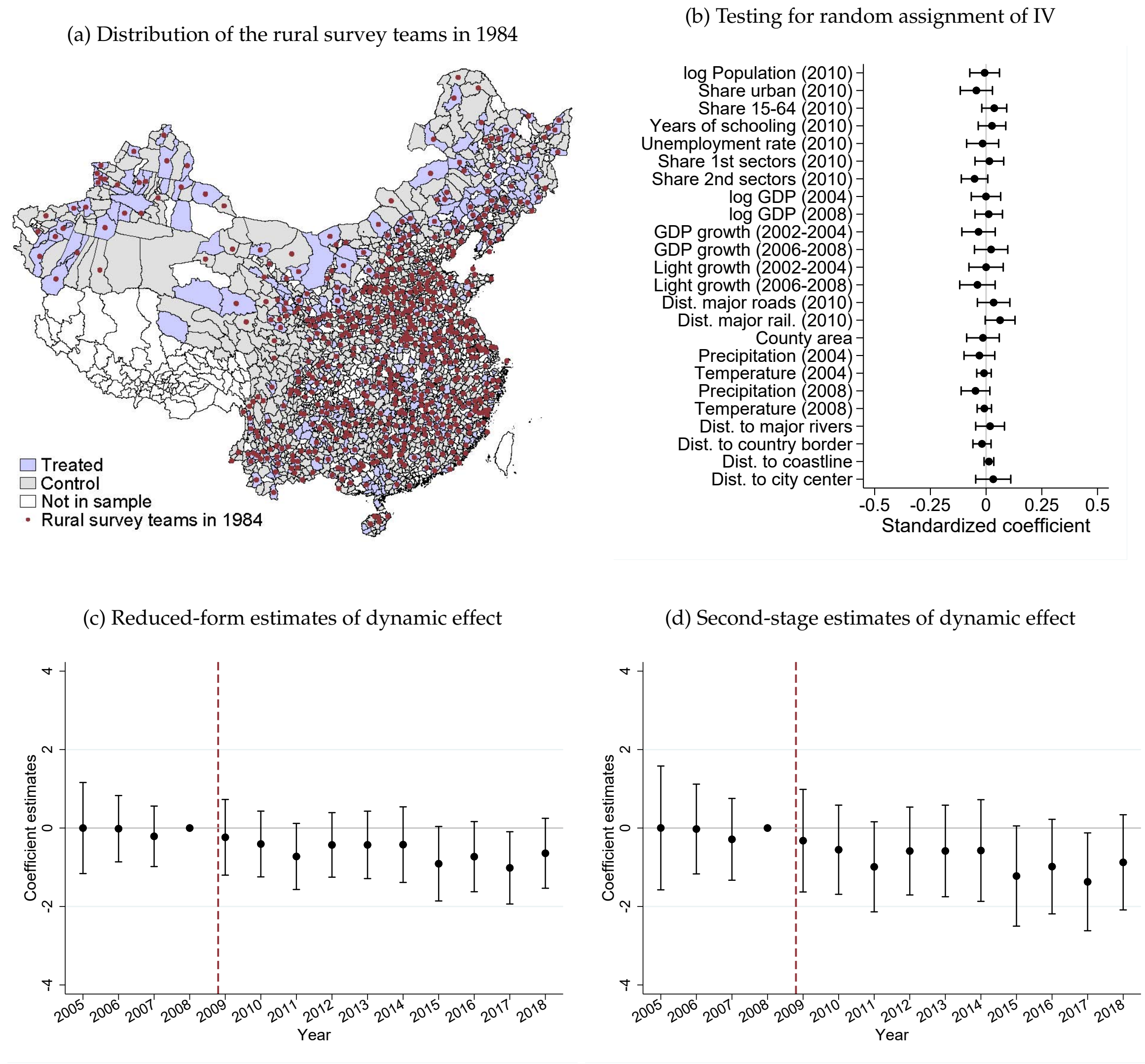


(b) Baseline + County covariates × Post



Notes: This figure conducts a placebo event study around 2005, when the survey teams were deployed but had not started disciplining local officials. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure 5: The randomly assigned rural survey teams in 1984 as IV

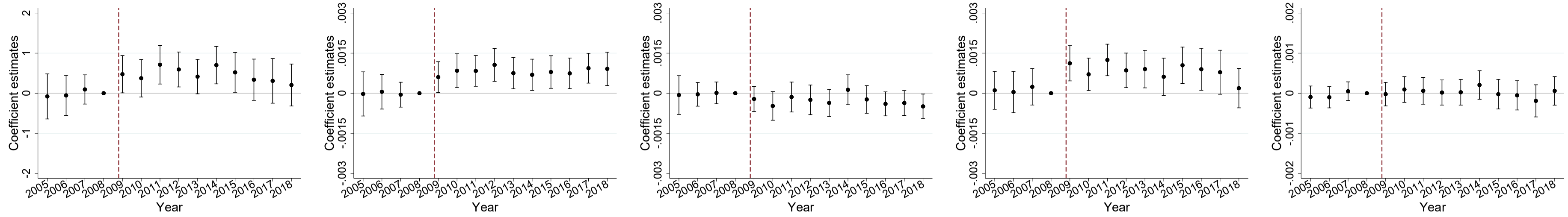


Notes: Panel (a) shows the distribution of the randomly assigned rural survey teams in 1984, which serves as an instrument for treatment counties. Panel (b) tests the randomness of the instrument and plots the standardized coefficients from a univariate regression of baseline county covariates on the instrument, with province fixed effects included as the randomization is stratified by province. Panel (c) shows the reduced-form event study estimated by replacing the treatment in the baseline event study specification (equation 6) with the instrument. Panel (d) shows the second-stage event study estimated by replacing the post-reform dummy in the baseline IV specification (equation 11) with a set of year dummies. The year 2008, which is one year before the reform in 2009, is omitted as the reference year. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure 6: Dynamic effect on government policies

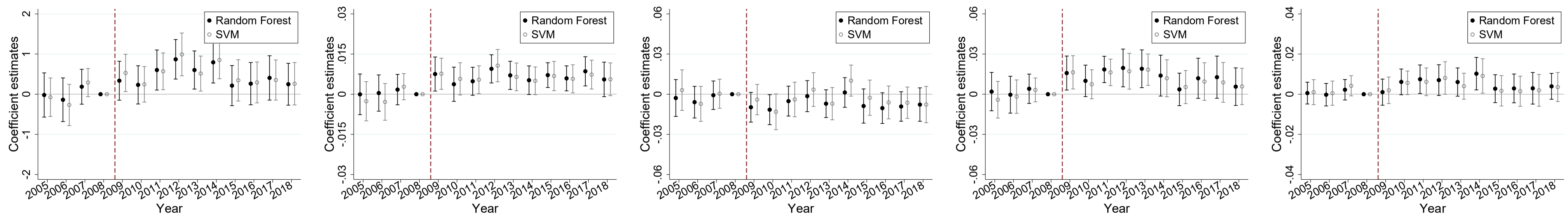
Key words frequency

(a) Standardized index (b) Business attraction (c) Infrastructure investment (d) Market reform (e) Policy experimentation



Machine learning-based topic score

(f) Standardized index (g) Business attraction (h) Infrastructure investment (i) Market reform (j) Policy experimentation

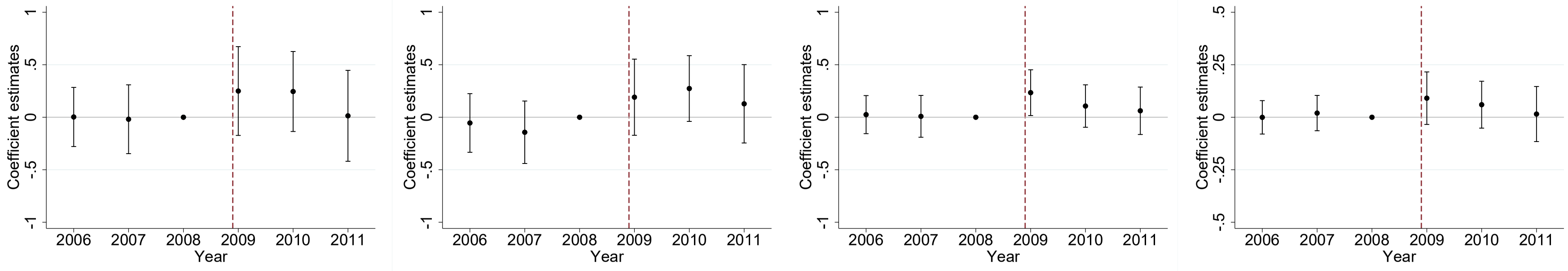


Notes: This figure shows the dynamic effect of the reform on government policies across four policy areas. The unit of observation is county. The sample period is 2005-2018. The sample includes 97 counties. The estimation equations are event study variants of the specifications in Table 5. The year 2008, one year before the reform in 2009, is omitted as the reference year. Panels (a)-(e) measure policies using a simple key words frequency method. Panels (f)-(j) measure policies using supervised machine learning methods (Random Forest and Support Vector Machine). The detailed procedures for constructing these measures are described in Appendix C. To alleviate multiple hypothesis testing issues, panels (a) and (f) report estimates using an standardized index by summarizing the four policy measures following Kling, Liebman and Katz (2007). Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure 7: Dynamic effect on bank credit

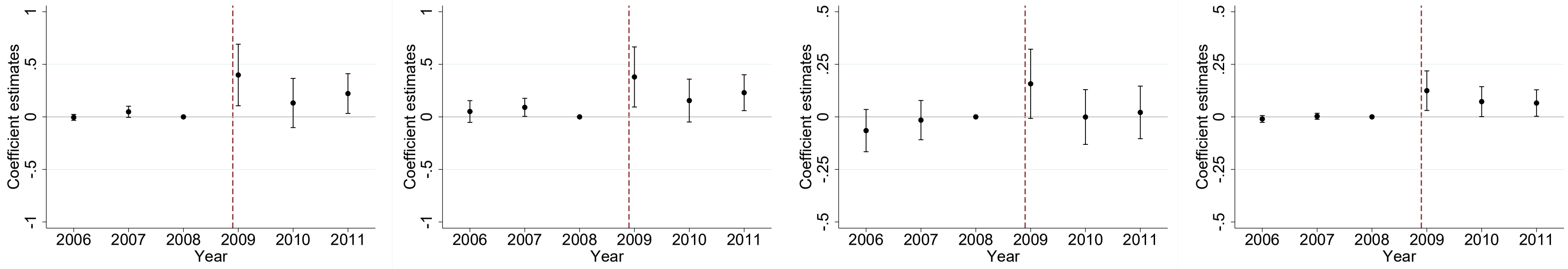
Difference-in-differences estimates

(a) IHS(Loans) (b) IHS(Loans to small firms) (c) IHS(# Firms granted bank loans) (d) IHS(# Branches with loan services)



Difference-in-difference-in-differences estimates

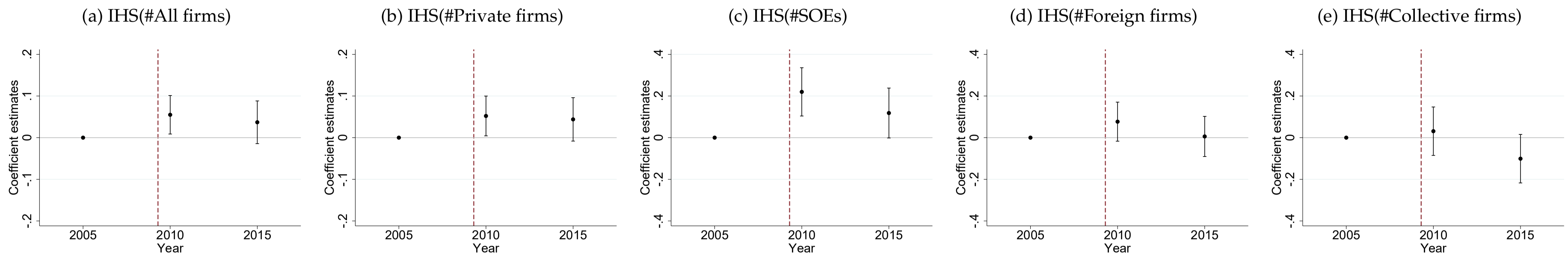
(e) IHS(Loans) (f) IHS(Loans to small firms) (g) IHS(# Firms granted bank loans) (h) IHS(# Branches with loan services)



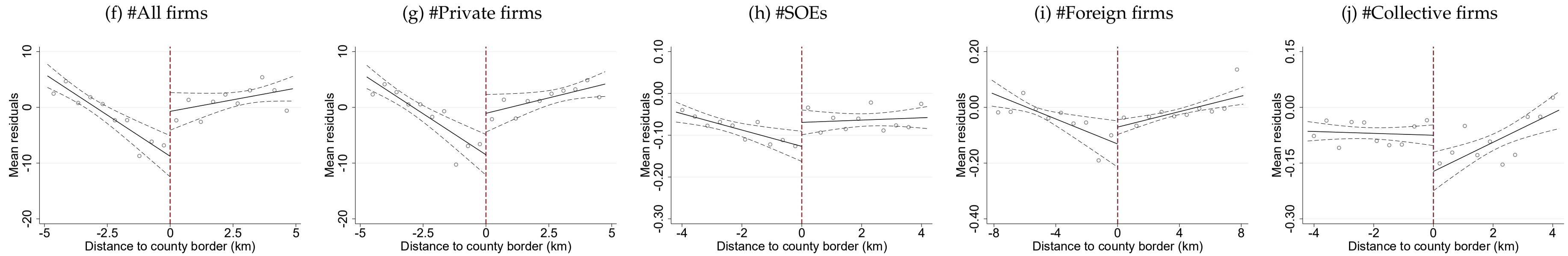
Notes: This figure shows the dynamic effect of the reform on bank credit. The unit of observation is county. The sample period is 2006-2011. The estimation equations are event study variants of the specifications in Table 6. The dependent variables are transformed by inverse hyperbolic sine (IHS) to reduce influences from the tails of the skewed outcome distributions (Bellemare and Wichman, 2020). Panels (a)-(d) conduct a conventional difference-in-differences estimation. Panels (e)-(h) further conduct a difference-in-difference-in-differences estimation utilizing differential control of banks by local governments across counties. Government control over banks is measured as the standardized share of pre-reform loans from City Commercial Banks (CCBs) in a county, whose controlling shareholders are local governments. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure 8: Effect on firm entry

County-level DID estimates



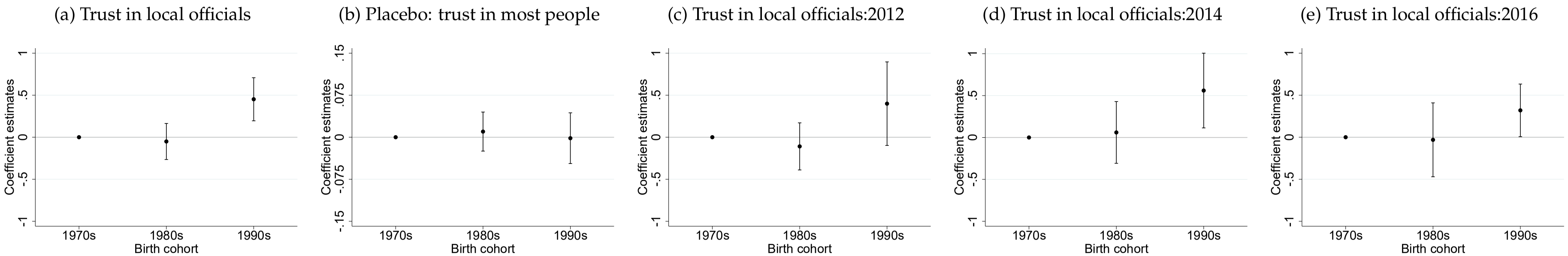
Town-level RD estimates



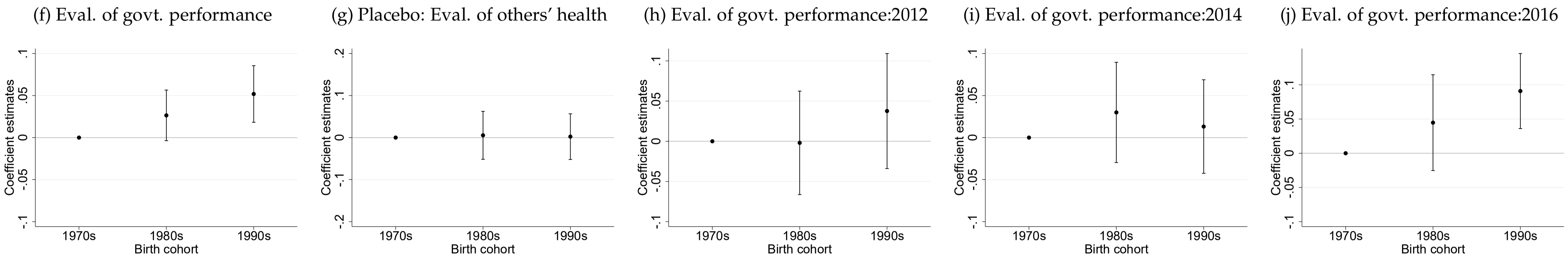
Notes: Panels (a)-(e) show the dynamic effect of the reform on firm entry at the county level using event study variants of the specifications in Table 7. The sample includes the years 2005, 2010, and 2015. The dependent variables are transformed by inverse hyperbolic sine (IHS) to reduce influences from the tails of the skewed outcome distributions (Bellemare and Wichman, 2020). Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level. Panels (f)-(j) show the jumps in firm entry at the town level across county borders with different treatment statuses on each side. The sample includes only the post-reform years 2010 and 2015. Due to the small number of firm entry at the town level, here the IHS transformations are not adopted (Bellemare and Wichman, 2020). The circles denote the mean of town-level firm entry within a distance bin, after partialling out county border fixed effects. The solid lines are fitted separately for each side of the county border. The dashes are the associated 90% confidence intervals constructed using standard errors clustered at the county border level. The bandwidths are the optimal bandwidths proposed by Calonico, Cattaneo and Titiunik (2014).

Figure 9: Effect on citizen attitudes towards local governments

Trust in local officials



Evaluation of government performance



Notes: This figure shows the effect of the reform on citizens' attitudes towards local governments, and is created by visualizing the results in Table 8. The unit of observation is citizen (three waves of survey data from the China Family Panel Studies, 2012, 2014 and 2016). The estimation equation is a cohort difference-in-differences specification (equation 14) that utilizes two sources of variation: (1) treatment counties versus control counties; (2) most affected cohorts versus less unaffected cohorts within the same county. The latter source of variation is built on insights in the psychology and political science literature that citizens' political attitudes are most permeable during teenage years and keep stable since one's 30s (Wolfinger and Rosenstone, 1980; Krosnick and Alwin, 1989). As such, the 1990s cohort are defined as the most affected group as they would be younger than 30 during the reform period, and thus, their political attitudes would be most permeable. The 1980s cohort are defined as the less affected group because some of them would be older than 30 during the reform period. The omitted group is those born in the 1970s, as they would be older than 30 in the reform period and thus be unaffected by the reform. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Table 1: Balance test

	Treat		Control		Difference		
	Mean	SD	Mean	SD	T - C	SE	*p*-value
			Panel A: Demography				
log Population (2010)	13.00	0.70	12.50	0.81	0.50	0.04	0.00
Share urban (%, 2010)	33.65	12.30	34.64	14.31	-0.99	0.66	0.13
Share 15-64 (%, 2010)	72.53	4.56	72.44	4.53	0.10	0.22	0.67
Years of schooling (2010)	8.23	0.73	8.19	1.05	0.04	0.05	0.34
			Panel B: Economic development				
Unemployment rate (%, 2010)	2.08	1.39	2.16	1.52	-0.09	0.07	0.23
Share primary sectors (%, 2010)	64.97	18.20	64.30	17.93	0.68	0.88	0.44
Share secondary sectors (%, 2010)	16.10	12.79	15.49	11.54	0.61	0.59	0.30
log GDP (2004)	12.47	0.94	11.97	1.03	0.50	0.05	0.00
log GDP (2008)	12.96	0.97	12.48	1.06	0.48	0.05	0.00
GDP growth (%, 2002-2004 average)	11.28	6.26	11.57	6.53	-0.29	0.33	0.38
GDP growth (%, 2006-2008 average)	12.86	6.38	12.67	6.63	0.19	0.32	0.54
Light growth (%, 2002-2004 average)	18.24	12.27	18.82	16.28	-0.58	0.72	0.42
Light growth (%, 2006-2008 average)	6.86	10.31	7.67	11.93	-0.81	0.55	0.14
Distance to major roads (km, 2010)	69.18	96.10	75.21	80.62	-6.03	4.24	0.16
Distance to major railways (km, 2010)	70.16	102.37	74.22	90.45	-4.06	4.63	0.38
			Panel C: Geography				
County area (km^2)	3900	7453	4128	10222	-228	446	0.61
Precipitation (inches, 2004)	0.04	0.09	0.03	0.10	0.00	0.00	0.55
Temperature (degrees, 2004)	13.91	5.23	13.57	5.41	0.34	0.26	0.19
Precipitation (inches, 2008)	0.05	0.11	0.05	0.11	0.01	0.01	0.20
Temperature (degrees, 2008)	13.73	5.08	13.43	5.25	0.30	0.25	0.23
Distance to major rivers (km)	59.17	61.15	57.76	59.03	1.41	2.90	0.63
Distance to country border (km)	346.52	251.00	345.25	251.49	1.26	12.16	0.92
Distance to coastline (km)	616.84	612.04	640.02	568.79	-23.18	28.46	0.42
Distance to prefecture center (km)	60.14	41.78	62.66	46.66	-2.52	2.17	0.25

Notes: This table provides balance tests by comparing counties with the survey teams deployed in 2005 to those without. The *p*-values reported in the last column are from *t*-tests of mean equality between groups. Except for GDP, all variables in this table are from sources that the county has no control on. The reason for using the year 2010 for the demographic data, instead of years before the reform in 2009, is that the population census was only conducted in 2000 and 2010. Similarly, the transportation data is also in 2010 due to data limitation

Table 2: Effect on GDP growth manipulation

	(1)	(2)	(3)	(4)
Dep. var.:	Reported GDP growth (%)			
Treat × Post	-0.751**	-0.869***	-0.552***	-0.576***
	(0.316)	(0.331)	(0.162)	(0.161)
Light growth (%)	0.023***	0.022***	0.017***	0.017***
	(0.005)	(0.005)	(0.005)	(0.005)
County FE	X	X	X	X
Year FE	X	X	X	X
Demographic controls × Post		X	X	X
Economic controls × Post			X	X
Geographic controls × Post				X
Cluster level	County	County	County	County
Observations	23,360	22,580	20,343	20,273
R-squared	0.269	0.269	0.362	0.362
Mean dep. var.	10.97	10.97	10.84	10.84

Notes: This table shows the effect of the reform on GDP growth manipulation. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Standard errors clustered at the county level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Estimating spillover effect

	(1)	(2)	(3)	(4)
Dep. var.:	Reported GDP growth (%)			
Treat × Post	-0.576***	-0.580***	-0.565***	-0.585***
	(0.161)	(0.164)	(0.164)	(0.162)
# Treat neighbors × Post		-0.010		
		(0.071)		
1(# Treat neighbors>0) × Post			0.108	
			(0.263)	
1(# Treat neighbors>Median=2) × Post				-0.133
				(0.178)
Light growth (%)	0.017***	0.017***	0.017***	0.017***
	(0.005)	(0.005)	(0.005)	(0.005)
County FE	X	X	X	X
Year FE	X	X	X	X
County controls × Post	X	X	X	X
Neighbor number FE × Post		X	X	X
Cluster level	County	County	County	County
Mean dep. var.	10.84	10.84	10.84	10.84
Mean number of neighbors	5.91	5.91	5.91	5.91
Mean number of treated neighbors	1.97	1.97	1.97	1.97

Notes: The tables shows the spillover effect of the reform on GDP growth manipulation. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. # Treat neighbors denote the number of treatment counties among a county's neighbors, where neighbors are defined as counties sharing a common boundary segment with a county. Standard errors clustered at the county level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Effect on GDP growth manipulation - IV estimates

	(1)	(2)	(3)	(4)
Panel A: First-stage estimates				
Dep. var.:	Treat × Post			
Treat[1984] × Post	0.739***	0.724***	0.727***	0.729***
	(0.017)	(0.015)	(0.016)	(0.016)
Light growth (%)	-0.000***	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Effective F-statistic	1,945	2,210	2,044	2,080
Critical value for 5% worst case bias	37.4	37.4	37.4	37.4
Panel B: Second-stage estimates				
Dep. var.:	Reported GDP growth (%)			
Treat × Post	-0.711*	-0.839**	-0.589***	-0.590***
	(0.400)	(0.402)	(0.205)	(0.203)
Light growth (%)	0.021***	0.021***	0.016***	0.017***
	(0.005)	(0.005)	(0.005)	(0.005)
Mean dep. var.	10.98	10.97	10.84	10.84
County FE	X	X	X	X
Year FE	X	X	X	X
Province FE × Post	X	X	X	X
Demographic controls × Post		X	X	X
Economic controls × Post			X	X
Geographic controls × Post				X
Cluster level	County	County	County	County
Observations	22,998	22,580	20,343	20,273

Notes: This table shows the IV estimates on the effect of the reform on GDP growth manipulation. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with a survey team deployed in 2005. Treat[1984] is a dummy variable indicating counties with a randomly assigned rural survey team in 1984. Post is a dummy variable indicating years after the reform in 2009. Panel A provides the first-stage estimates. Panel B provides the second-stage estimates with Treat[1984] serving as an instrument for Treat. The effective F-statistics and corresponding critical values are constructed following Montiel Olea and Pflueger (2013). Standard errors clustered at the county level are reported in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 5: Effect on government policies

Dep. var.:	(1) Standardized index	(2) Business attraction	(3) Infrastructure investment	(4) Market reform	(5) Policy experimentation
Panel A: Key words frequency					
Treat × Post	0.474***	0.001***	-0.000	0.001***	0.000
	(0.167)	(0.000)	(0.000)	(0.000)	(0.000)
R-squared	0.482	0.513	0.481	0.606	0.615
Mean dep. var.	0	0.002	0.002	0.004	0.001
Panel B: Topic score predicted by Random Forest					
Treat × Post	0.452***	0.006***	-0.005	0.012**	0.004*
	(0.155)	(0.002)	(0.004)	(0.005)	(0.002)
R-squared	0.427	0.435	0.361	0.591	0.512
Mean dep. var.	0	0.013	0.040	0.068	0.016
Panel C: Topic score predicted by Support Vector Machine					
Treat × Post	0.508***	0.007***	-0.003	0.012***	0.003
	(0.163)	(0.002)	(0.005)	(0.004)	(0.002)
R-squared	0.474	0.427	0.435	0.613	0.587
Mean dep. var.	0	0.014	0.051	0.068	0.019
County FE	X	X	X	X	X
Year FE	X	X	X	X	X
County controls × Post	X	X	X	X	X
Cluster level	County	County	County	County	County
Observations	883	883	883	883	883

Notes: This table shows the effect of the reform on government policies across four policy areas. The unit of observation is county. The sample period is 2005-2018. The sample includes 97 counties. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Panel A measures policies using a simple key words frequency method. Panel B and C measure policies using supervised machine learning methods (Random Forest and Support Vector Machine). The detailed procedures for constructing these measures are described in Appendix C. To alleviate multiple hypothesis testing issues, column (1) reports estimates using an standardized index by summarizing the four policy measures following Kling, Liebman and Katz (2007). Standard errors clustered at the county level are reported in parentheses. * p <0.1, ** p <0.05, *** p <0.01.

Table 6: Effect on bank credit

Dep. var. is IHS of:	(1) Total amount of loans	(2) Loans to small firms	(3) # Firms granted loans	(4) # Branches granting loans
Panel A: Difference-in-differences				
Treat × Post	0.174	0.263**	0.123	0.049
	(0.123)	(0.116)	(0.076)	(0.036)
R-squared	0.329	0.424	0.520	0.482
Panel B: Difference-in-difference-in-differences				
Treat × Government control × Post	0.237**	0.208**	0.086	0.090***
	(0.103)	(0.097)	(0.062)	(0.033)
Treat × Post	0.027	0.062	0.024	-0.009
	(0.124)	(0.117)	(0.073)	(0.036)
Government control × Post	0.026	0.026	0.062	0.017
	(0.089)	(0.081)	(0.050)	(0.029)
R-squared	0.325	0.424	0.551	0.514
County controls × Post	X	X	X	X
County FE	X	X	X	X
Year FE	X	X	X	X
Observations	8,922	8,922	8,922	8,922

Notes: This table shows the effect of the reform on bank credit. The unit of observation is county. The sample period is 2006-2011. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Government control denotes the standardized share of pre-reform loans from City Commercial Banks (CCBs) in a county, whose controlling shareholders are local governments. The dependent variables are transformed by inverse hyperbolic sine (IHS) to reduce influences from the tails of the skewed outcome distributions (Bellemare and Wichman, 2020). Panel A conducts a difference-in-differences estimation as usual. Panel B adopts a difference-in-difference-in-differences estimation utilizing differential control of banks by local governments across counties. The estimation equation is equation (12). Standard errors clustered at the county level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Effect on firm entry

	(1)	(2)	(3)	(4)	(5)
Firm type:	All	Private	SOEs	Foreign	Collective
Panel A: Difference-in-differences estimates					
Dep. var.:	IHS(# Firm registrations)				
Treat × Post	0.046*	0.048*	0.169***	0.041	-0.035
	(0.026)	(0.027)	(0.062)	(0.051)	(0.062)
County FE	X	X	X	X	X
Year FE	X	X	X	X	X
County controls × Post	X	X	X	X	X
Cluster level	County	County	County	County	County
Observations	4,494	4,494	4,494	4,494	4,494
R-squared	0.943	0.944	0.535	0.726	0.571
Panel B: Regression discontinuity estimates at the town level					
Dep. var.:	# Firm registrations in post-reform period				
Treat	8.594**	8.241**	0.035	0.016	-0.084**
	(3.378)	(3.369)	(0.027)	(0.025)	(0.037)
County border FE	X	X	X	X	X
Cluster level	County border	County border	County border	County border	County border
RD polynomial	Linear	Linear	Linear	Linear	Linear
Observations	10,776	10,492	9,266	15,356	9,304
R-squared	0.334	0.335	0.229	0.372	0.249
Mean dep. var.	18.50	18.35	0.10	0.09	0.06
Bandwidth (in km)	4.88	4.74	4.20	8.14	4.21

Notes: This table shows the effect of the reform on firm entry. The unit of observation is county in panel A and town in panel B. The sample includes the years 2005, 2010, and 2015 in panel A and the years 2010 and 2015 (post-reform years) in panel B. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Panel A conducts a difference-in-differences estimation as usual. Panel B conducts a regression discontinuity estimation at the town level along county borders with different treatment statuses on each side. The estimation equation is a local linear specification (equation 13) using the optimal bandwidth proposed by Calonico, Cattaneo and Titiunik (2014). The dependent variables at the county level in panel A are transformed by inverse hyperbolic sine (IHS) to reduce influences from the tails of the skewed outcome distributions (Bellemare and Wichman, 2020). For the dependent variables at the town level in panel B, such transformations are not adopted due to the small number of firm entry (Bellemare and Wichman, 2020). Standard errors clustered at the indicated (county or county border) level are reported in parentheses. * p <0.1, ** p <0.05, *** p <0.01.

Table 8: Effect on citizens attitudes towards local governments

	(1)	(2)	(3)	(4)	(5)
Panel A: Trust in local officials					
Dep. var.:	Trust in local officials: 2012-2016	Placebo: trust in most people: 2012-2016	Trust in local officials: 2012	Trust in local officials: 2014	Trust in local officials: 2016
Treat × 1980s cohort	-0.051	0.010	-0.109	0.060	-0.031
	(0.129)	(0.021)	(0.168)	(0.221)	(0.264)
Treat × 1990s cohort	0.452***	-0.002	0.399	0.560**	0.319*
	(0.154)	(0.027)	(0.299)	(0.268)	(0.188)
Observations	17,163	17,166	6,338	5,651	5,174
R-squared	0.069	0.086	0.096	0.110	0.089
Number of counties	82	82	81	82	82
Mean dep. var.	4.84	0.57	4.82	4.90	4.80
Panel B: Evaluation of local government performance					
Dep. var.:	Eval. of govt. performance: 2012-2016	Placebo: eval. of others' health: 2012-2016	Eval. of govt. performance: 2012	Eval. of govt. performance: 2014	Eval. of govt. performance: 2016
Treat × 1980s cohort	0.026	0.004	-0.002	0.030	0.045
	(0.018)	(0.034)	(0.039)	(0.036)	(0.042)
Treat × 1990s cohort	0.052**	0.002	0.038	0.013	0.091***
	(0.020)	(0.033)	(0.043)	(0.033)	(0.033)
Observations	17,194	11,999	6,353	5,659	5,182
R-squared	0.060	0.108	0.097	0.096	0.081
Number of counties	82	82	81	82	82
Mean dep. var.	0.82	0.25	0.53	0.81	0.85
County FE	X	X	X	X	X
Cohort FE	X	X	X	X	X
County controls × Cohort FE	X	X	X	X	X
Survey wave FE	X	X	X	X	X
Citizen controls	X	X	X	X	X
Cluster level	County	County	County	County	County

Notes: This table shows the effect of the reform on citizens' attitudes towards local governments. The unit of observation is citizen (three waves of survey data from the China Family Panel Studies, 2012, 2014 and 2016). Treat is a dummy variable indicating counties with the survey teams deployed in 2005. 1980s cohort and 1990s cohort are dummy variables indicating citizens born in the 1980s and the 1990s, respectively. The estimation equation is a cohort difference-in-differences specification (equation 14) that utilizes two sources of variation: (1) treatment counties versus control counties; (2) most affected cohorts versus less unaffected cohorts within the same county. The latter source of variation is built on insights in the psychology and political science literature that citizens' political attitudes are most permeable during teenage years and keep stable since one's 30s (Wolfinger and Rosenstone, 1980; Krosnick and Alwin, 1989). As such, the 1990s cohort are defined as the most affected group as they would be younger than 30 during the reform period, and thus, their political attitudes would be most permeable. The 1980s cohort are defined as the less affected group because some of them would be older than 30 during the reform period. The omitted group is those born in the 1970s, as they would be older than 30 in the reform period and thus be unaffected by the reform. Standard errors clustered at the county level are reported in parentheses. * p <0.1, ** p <0.05, *** p <0.01.

Online Appendix

Curbing Bureaucratic Information Manipulation

Yongwei Nian

A Theory appendix

This section shows how to solve for the equilibrium effort m^* (effort in manipulating GDP growth) and e^* (effort in stimulating the economy). To write leader i's maximization problem, note that there are four situations: (1) with probability $(1-p)^2$, both i and $-i$ are not detected for manipulation; (2) with probability $(1-p)p$, i is not detected for manipulation, but $-i$ is; (3) with probability $p(1-p)$, i is detected for manipulation, but $-i$ is not; (4) with probability p^2, both are detected for manipulation. Let q_1, q_2, q_3, q_4 denote i's promotion probability in each case, respectively:

$$\begin{aligned} q_1 &= Pr(G_i + \varepsilon_i > G_{-i} + \varepsilon_{-i}) \\ &= \frac{1}{2} + \phi[h(m_i) + g(e_i) - h(m_{-i}) - g(e_{-i})] \end{aligned}$$

$$\begin{aligned} q_2 &= Pr[G_i + \varepsilon_i > G_{-i} + \varepsilon_{-i} - \delta h(m_{-i})] \\ &= \frac{1}{2} + \phi[h(m_i) + g(e_i) - (1-\delta)h(m_{-i}) - g(e_{-i})] \end{aligned}$$

$$\begin{aligned} q_3 &= Pr[G_i + \varepsilon_i - \delta h(m_i) > G_{-i} + \varepsilon_{-i}] \\ &= \frac{1}{2} + \phi[(1-\delta)h(m_i) + g(e_i) - h(m_{-i}) - g(e_{-i})] \end{aligned}$$

$$\begin{aligned} q_4 &= Pr[G_i + \varepsilon_i - \delta h(m_i) > G_{-i} + \varepsilon_{-i} - \delta h(m_{-i})] \\ &= \frac{1}{2} + \phi[(1-\delta)h(m_i) + g(e_i) - (1-\delta)h(m_{-i}) - g(e_{-i})] \end{aligned}$$

Then one can write i's expected payoff Z as:

$$\begin{aligned} Z = &(1-p)^2 q_1 u(R) + (1-p)^2(1-q_1)u(r) \\ &+ (1-p)pq_2 u(R) + (1-p)p(1-q_2)u(r) \\ &+ p(1-p)q_3[u(R) - \lambda h(m_i)] + p(1-p)(1-q_3)[u(r) - \lambda h(m_i)] \\ &+ p^2 q_4[u(R) - \lambda h(m_i)] + p^2(1-q_4)[u(r) - \lambda h(m_i)] \end{aligned}$$

Substituting q_1, q_2, q_3, q_4 into Z, i' problem can be write as:

$$\max_{e_i, m_i, e_i+m_i<=\bar{C}} u(r) + [u(R) - u(r)]\phi g(e_i) + [(u(R) - u(r))\phi(1 - p\delta) - \lambda p]h(m_i)$$

Assume $(u(R) - u(r))\phi(1 - p\delta) - \lambda p > 0$ and interior solution, one can write the first-order condition as:

$$[u(R) - u(r)]\phi g'(e_i^*) = [(u(R) - u(r))\phi(1 - p\delta) - \lambda p]h'(m_i^*)$$

Also note that at equilibrium $m_i^* + e_i^* = \bar{C}$. Let $K(\cdot)$=$[\frac{h'(\cdot)}{g'(c-\cdot)}]^{-1}$, then one could solve for m_i^* and e_i^*:

$$m^* = K\left[\frac{V\phi}{V\phi(1 - p\delta) - \lambda p}\right]$$
$$e^* = \bar{C} - m^*$$

where $V = u(R) - u(r)$.

B Characterizing compliers in IV estimation

In this section, I follow Marbach and Hangartner (2020) to understand the characteristics of compliers in the presence of heterogeneous treatment effects in the IV estimation. Let *Treat* denote the realized treatment status of all counties, which is equal to 1 if a county had survey team deployed in 2005, and 0 otherwise. Let $Treat^{1984}$ denote the instrument, which is equal to 1 if a county had a randomly assigned rural survey team in 1984, and 0 otherwise. Let $Treat(0)$ and $Treat(1)$ denote the potential treatment status of a county depending on the instrument. Under monotonicity, which is supported by Appendix Table A3, defier counties can be ruled out. Then under random assignment of the instrument, which is also true as I discuss in Section 4.2.3, the mean of covariate X among never-taker counties could be computed using their observed counterparts:

$$E[X|NeverTakers] = E[X|Treat = 0, Treat^{1984} = 1]$$

The covariate mean for always-taker counties could also be computed using their observed counterparts:

$$E[X|AlwaysTakers] = E[X|Treat = 1, Treat^{1984} = 0]$$

Under random assignment of the instrument, one could also use the realized values of *Treat* to compute the fraction of never-taker counties and always-taker counties as follows, respectively:

$$Pr[NeverTaker] = Pr[Treat = 0|Treat^{1984} = 1] = 1 - E[Treat|Treat^{1984} = 1]$$

$$Pr[AlwaysTaker] = Pr[Treat = 1|Treat^{1984} = 0] = E[Treat|Treat^{1984} = 0]$$

Under monotonicity, the fraction of complier counties is equal to:

$$Pr[Complier] = 1 - Pr[NeverTaker] - Pr[AlwaysTaker]$$

Then according to the law of total expectation, the mean of covariate X can be written as:

$$\begin{aligned} E[X] = & E[X|Complier]Pr[Complier] \\ & + E[X|NeverTaker]Pr[NeverTaker] \\ & + E[X|AlwaysTaker]Pr[AlwaysTaker] \end{aligned}$$

Combining all these equations, one can back out the mean of covariate X among compliers, which takes the following form and can be empirically estimated:

$$\begin{aligned} E[X|Complier] = & \frac{E[X] - E[X|Treat = 0, Treat^{1984} = 1][1 - E[Treat|Treat^{1984} = 1]]}{E[Treat|Treat^{1984} = 1] - E[Treat|Treat^{1984} = 0]} \\ & - \frac{E[X|Treat = 1, Treat^{1984} = 0]E[Treat|Treat^{1984} = 0]}{E[Treat|Treat^{1984} = 1] - E[Treat|Treat^{1984} = 0]} \end{aligned}$$

C Textual analysis of government work reports

To construct textual measures of the government policies mentioned in Section 5.1, I first convert all county government work reports into 326,435 sentences, of which 25% are selected at random and manually labeled as belonging to one policy area or not. I then remove unnecessary characters (spaces, numbers, and punctuations) and stopwords from each sentence using a commonly used list of such characters and words.[1] As Chinese words are not space-delimited, I next adopt the open-source Chinese text segmentation library called *Jieba* to segment each sentence into words.[2] For the first textual measure, namely the keywords frequency measure, I simply count in a report the total number of mentions of the keywords corresponding to a certain policy (see Appendix Table A12 for the list of key words) and then divide this count by the total number of words in the report.

[1]The list is from https://github.com/goto456/stopwords/blob/master/README.md

[2]The library can be found at https://github.com/fxsjy/jieba

Then to construct the machine learning-based textual measures, I convert the tokenized sentences into numerical vectors using the term frequency-inverse document frequency (TF-IDF) method. This method weights each term (word) with its frequency in a document (sentence) and its inverse document (sentence) frequency.[3] So words with higher TF-IDF scores are of higher informativeness. I adopt two commonly used machine learning algorithms: random forest (RF) and support vector machine (SVM) (Gentzkow, Kelly and Taddy, 2019). The RF algorithm creates a forest of decision trees, with each tree trained on a random subset of the vectorized data points. To classify a new sentence, each tree in the forest can make a decision and the decision receiving the most votes is chosen as the final classification. The SVM algorithm operates by finding the optimal hyperplane that separates the two classess of vectorized data points. The optimization is acheived by maximizing the distance between the hyperplane and the cloest data points from either class. A new sentence is then vectorized and assigned to a class based on which side of the hyperplane it falls on. The outputs from these two algorithms are sentence-level binary policy scores and I take the sentence length weighted average of these scores to get a report-level policy score.

Finally, to compute textual similarity between two reports as used in Section 5.5, I use the aforementioned TF-IDF method to vectorize an entire report, and then calculate the pairwise cosine similarity between any two reports i and j following Kelly et al. (2021):

$$Similarity_{ij} = \frac{V_i \cdot V_j}{\parallel V_i \parallel \cdot \parallel V_j \parallel}$$

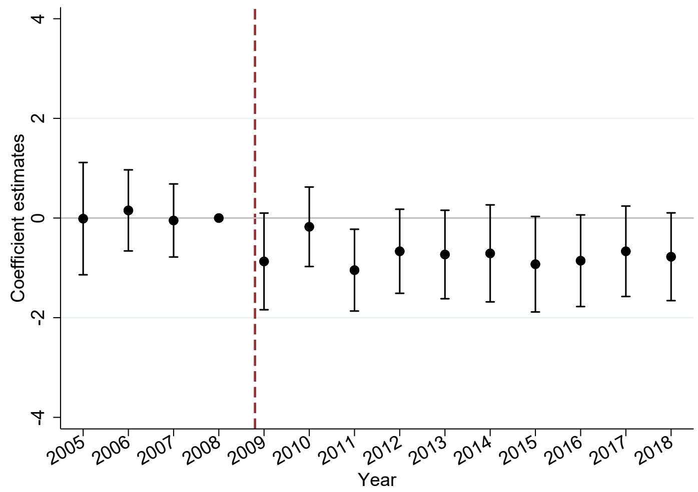
where V_i and V_j are vectors representig the two reports, and $V_i \cdot V_j$ is the dot product of these two vectors. $\parallel V_i \parallel$ and $\parallel V_j \parallel$ are norms of these two vectors. The output $Similarity_{ij}$ is then a measure between -1 and 1, with higher values denoting higher similarity.

[3] The IDF of a word in my case is defined as log(# sentences in sample / # sentences containing the word).

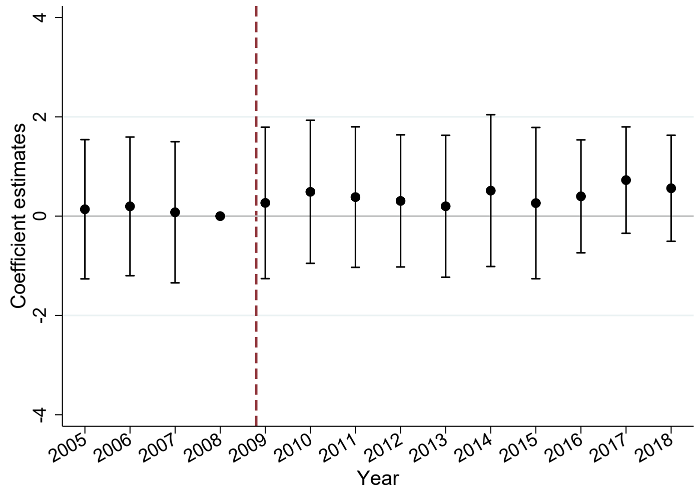
D Additional figures

Figure A1: Decomposing the effect of the reform

(a) Effect on reported GDP growth



(b) Effect on light growth

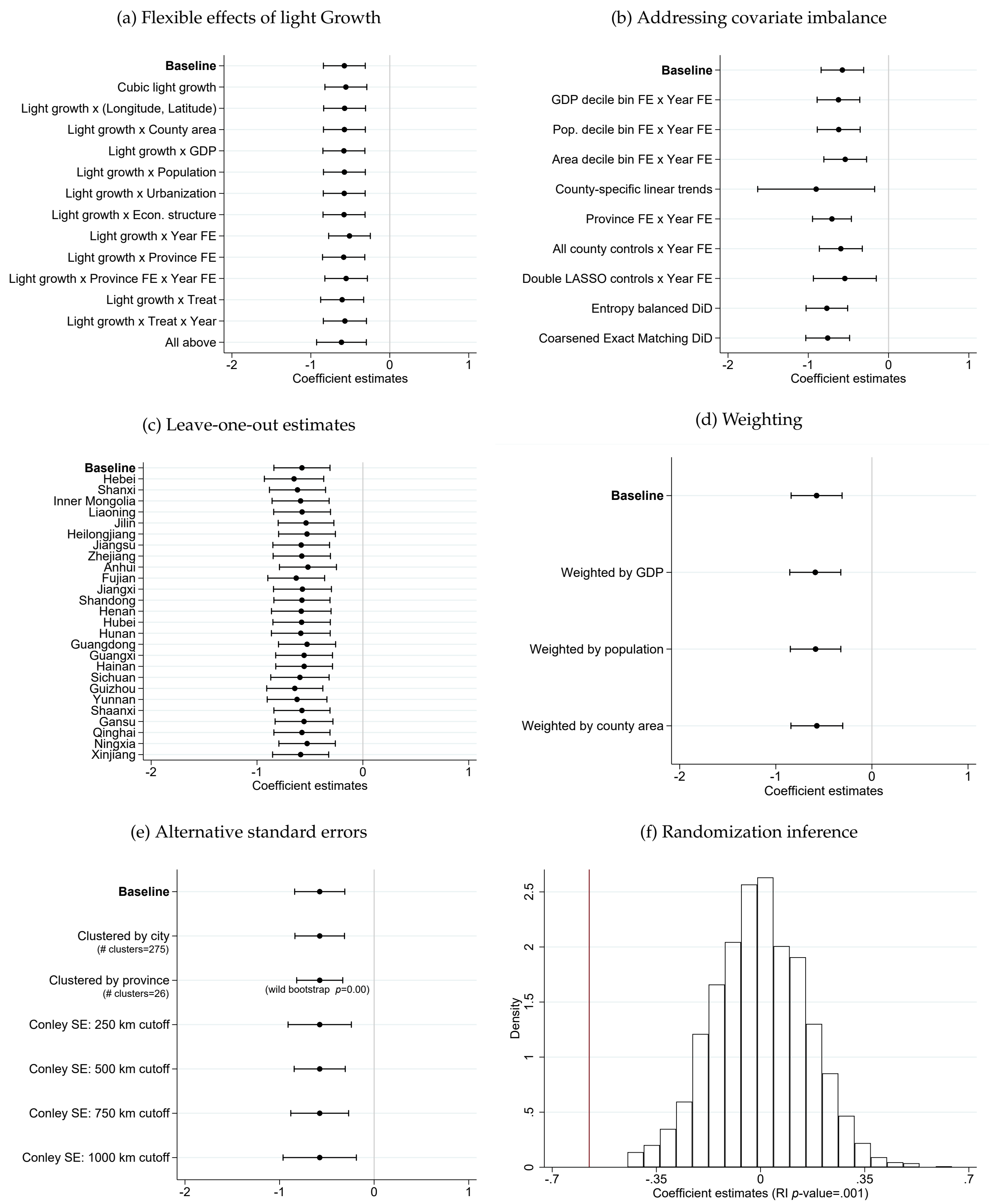


Notes: This figure shows the dynamic effect of the reform on reported GDP growth (panel a) and light growth (panel b) separately. Specifically, I estimate:

$$Y_{ct} = \sum_{j=2005,\, j\neq 2008}^{j=2018} \beta_j Treat_c \times 1_{\{t=j\}} + \delta_c + \lambda_t + \varepsilon_{ct}$$

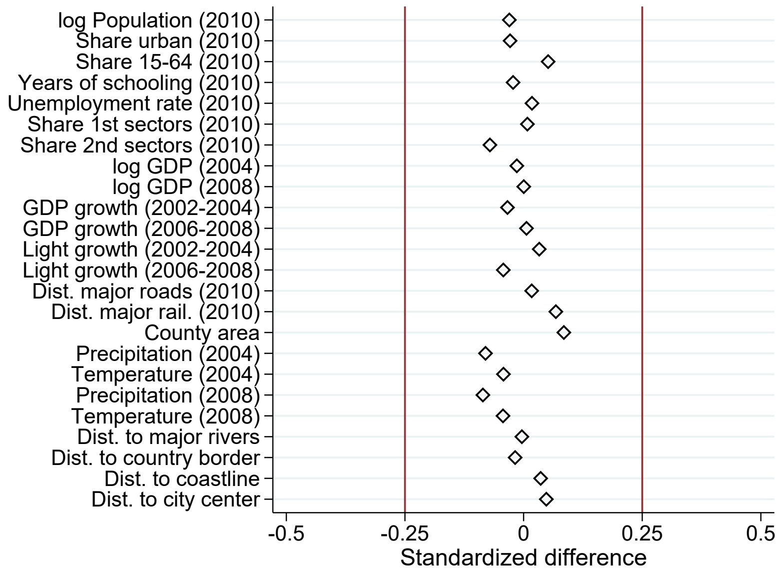
where Y_{ct} denotes either reported GDP growth (panel a) or light growth (panel b). Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure A2: Additional Robustness Checks on Baseline Estimates



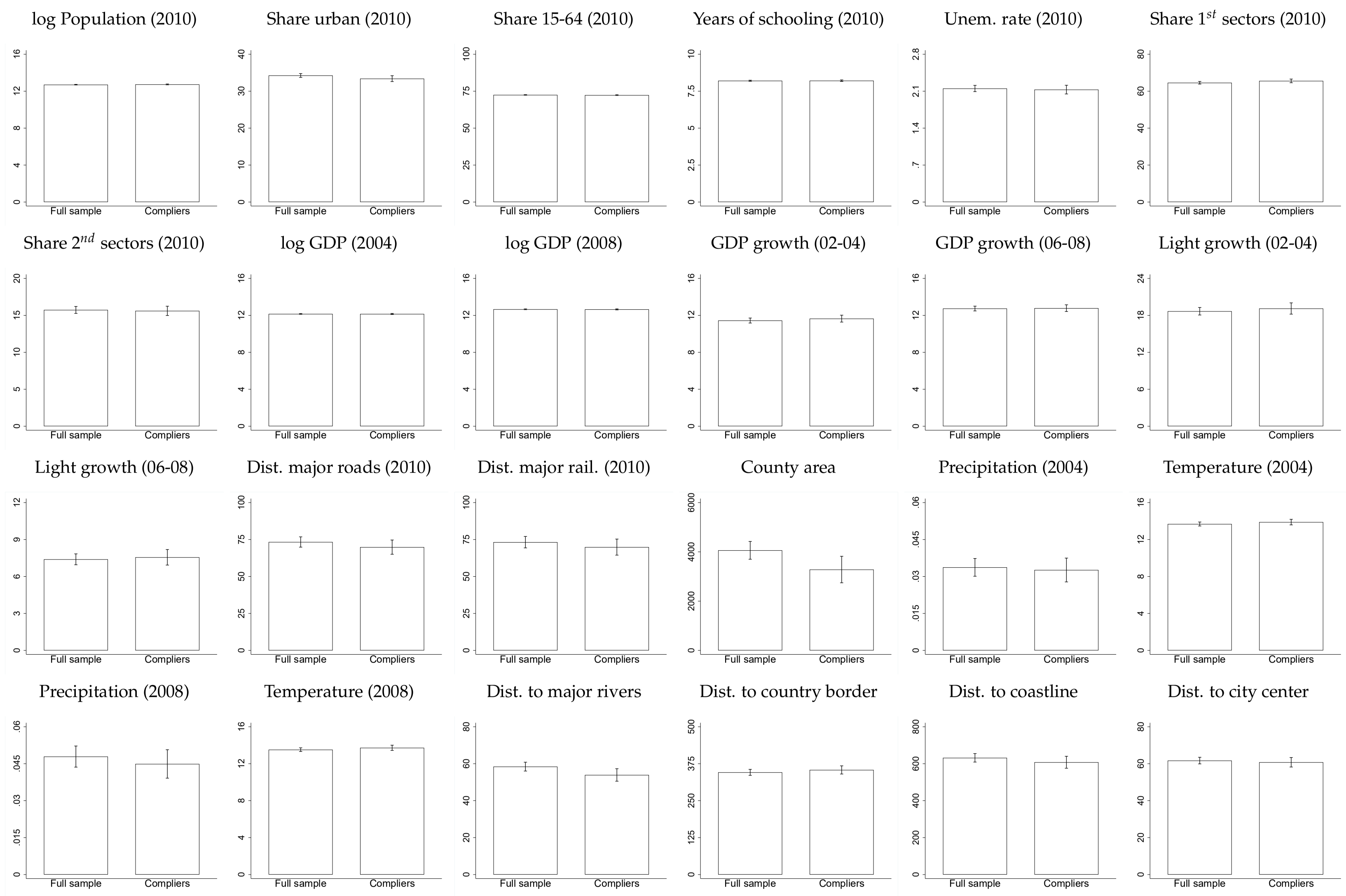
Notes: This figure provides additional robustness checks to the baseline estimates. All panels control for baseline county covariates interacted the post-reform dummy when appropriate. Panel (a) addresses the concern that the mapping between light growth and economic growth may not be uniform across counties or years, by allowing the mapping to vary flexibly. Panel (b) addresses the concern that the results may be confounded by covariate imbalance between treatment counties and control counties, by directly controlling for the sources of imbalance or achieving covariate balance through entropy balance and coarsened exact matching, among others. Panel (c) shows the estimates after excluding one province each time. Panel (d) shows the estimates weighted by baseline county size. Panel (e) shows the estimates with alternative clustering methods and spatial correlation correction. For province-level clustering with a small number of clusters of 26, I also report the wild bootstrap p-values with 2,000 replications (Roodman et al., 2019). For Conley standard errors (Conley, 1999), I account for serial correlation spanning all years and spatial correlation within distances of 250 km, 500 km, 750 km, and 1,000 km. Panel (f) shows the estimates using a randomization inference procedure with 2,000 permutations following Young (2019). The true estimate is denoted by the vertical line and the randomization inference p-value is reported below the figure. Standard errors used to construct the 90% confidence intervals, which are denoted by the spikes, are clustered at the county level when appropriate.

Figure A3: Standardized differences for the instrument



Notes: This graph shows the standardized differences between county groups defined by the instrument ($Treat^{1984} = 1$ vs $Treat^{1984} = 0$). To this end, I first compute the standardized differences between the two groups within each province, which are differences between the sample means normalized by the square root of the average of the sample variances. I then calculate a weighted average using the number of counties within each province as weights. The two vertical lines denote the 25% threshold recommended by Imbens and Rubin (2015) for covariate balance.

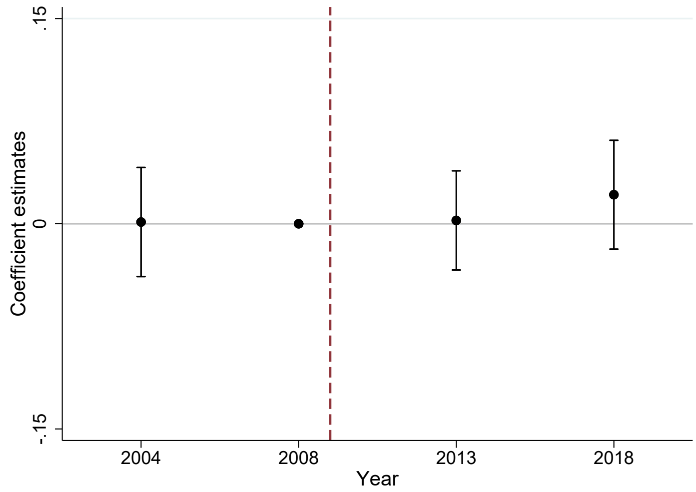
Figure A4: Complier characteristics



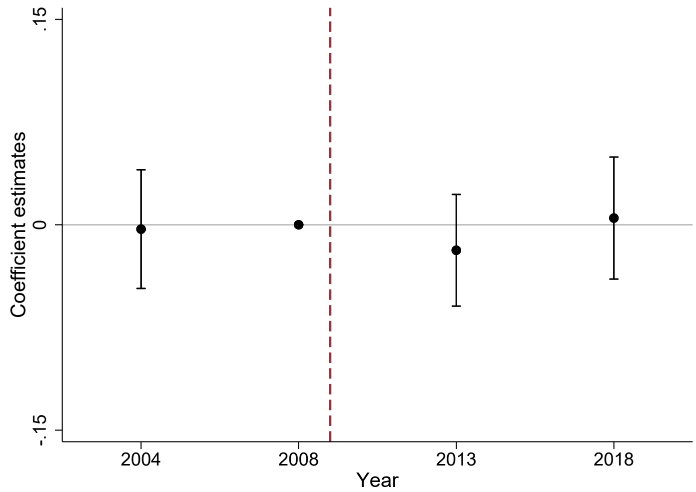
Notes: This figure plots the complier and full sample means, where the complier means are calculated following Marbach and Hangartner (2020) (see also Appendix B)

Figure A5: Dynamic effect on local statistical capacity

(a) Baseline specification



(b) Baseline + County covariates × Post

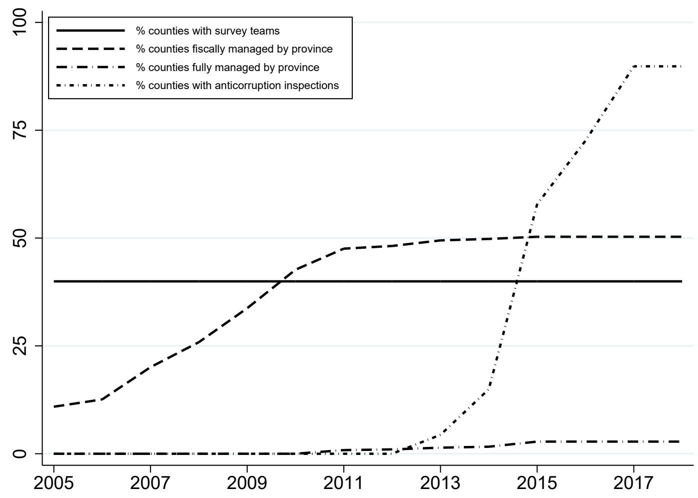


Notes: This figure shows the dynamic effect of the reform on local statistical capacity, and is estimated using the following specification:

$$Y_{ct} = \sum_{k=2004,2013,2018,k\neq 2008} \beta_k Treat_c \times 1_{\{t=k\}} + \delta_c + \lambda_g + \varepsilon_{ct} \tag{A1}$$

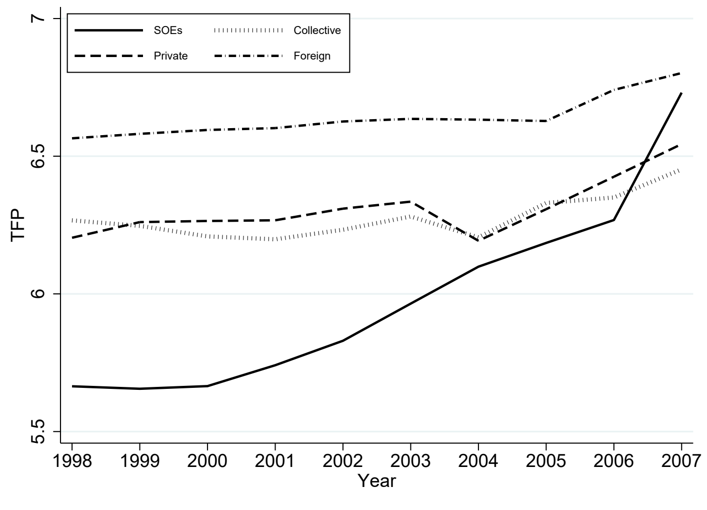
where Y_{ct} is a dummy variable equal to 1 if county c won an award for outstanding performance in conducting economic census in year t. The year 2008 is omitted as the reference group. This award is used as a proxy for local statistical capacity. The data on this award is only available for 2004, 2008, 2013, and 2018. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure A6: Rollout of concurrent reforms



Notes: This figure shows the rollout of various concurrent reforms. They include the fiscal province-managing-county (PMC) reform, the full province-managing-county (PMC) reform, and the anti-corruption campaign launched in 2013.

Figure A7: TFP by ownership

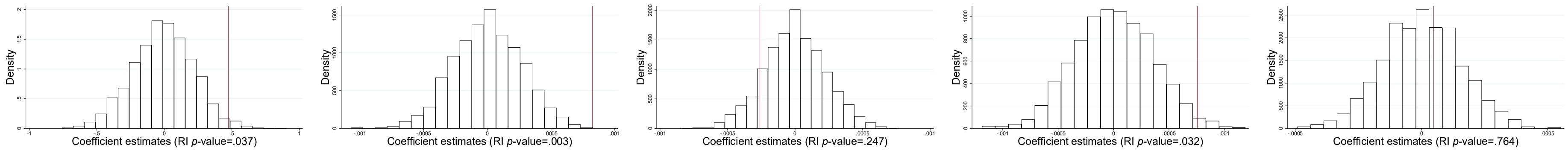


Notes: This figure shows the dynamics of firm total factor productivity (TFP) by ownership. It is created by taking the average TFP by ownership for manufacturing firms from 1998 to 2007 using the commonly used Annual Survery of Industrial Firms dataset. The calculation of the TFP uses the Levinsohn and Petrin method (Levinsohn and Petrin, 2003).

Figure A8: Effect on government policies - randomization inference

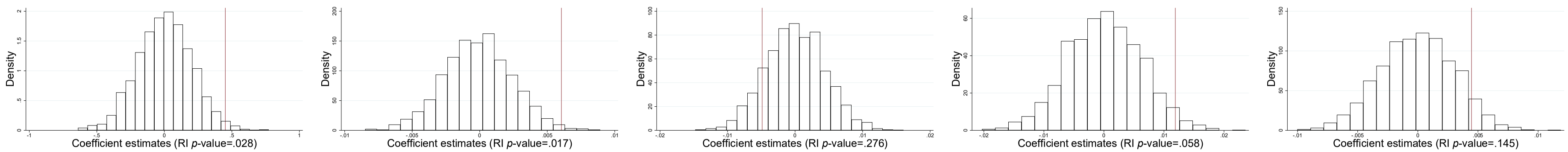
Key words frequency

(a) Standardized index (b) Business attraction (c) Infrastructure investment (d) Market reform (e) Policy experimentation



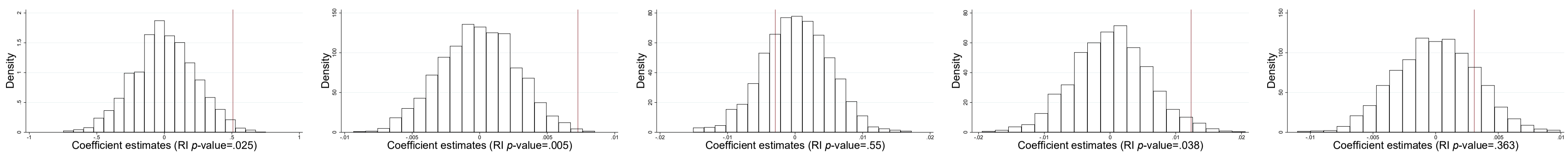
Topic score predicted by Random Forest

(f) Standardized index (g) Business attraction (h) Infrastructure investment (i) Market reform (j) Policy experimentation



Topic score predicted by Support Vector Machine

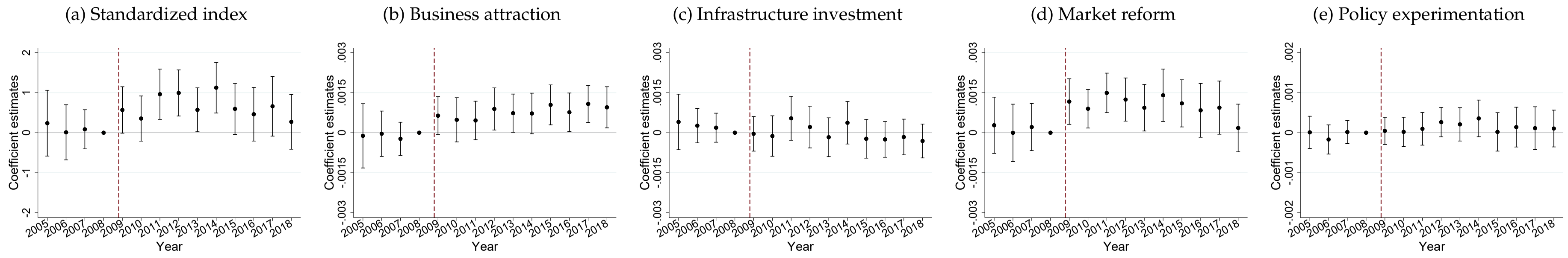
(k) Aggregate index (l) Business attraction (m) Infrastructure investment (n) Market reform (o) Policy experimentation



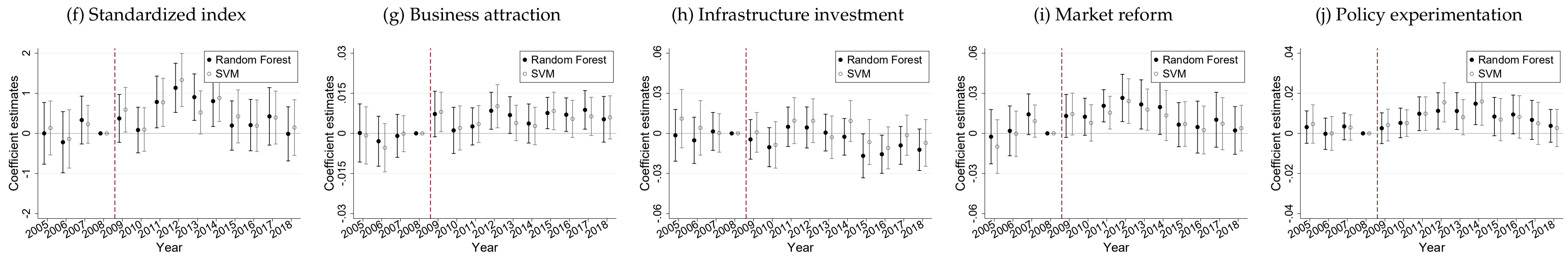
Notes: This figure adopts a randomization inference procedure with 2,000 permutations following Young (2019) to alleviate concerns about small sample size when estimating the effect of the reform on government policies. The true estimates are denoted by the vertical lines and the randomization inference p-values are reported below the figures. The unit of observation is county. The sample period is 2005-2018. The sample includes 97 counties. Panels (a)-(e) measure policies using a simple key words frequency method. Panels (f)-(o) measure policies using supervised machine learning methods (Random Forest and Support Vector Machine). The detailed procedures for constructing these measures are described in Appendix C. To alleviate multiple hypothesis testing issues, panels (a), (f), and (k) report estimates using a standardized index by summarizing the four policy measures following Kling, Liebman and Katz (2007).

Figure A9: Dynamic effect on government policies - IV estimates

Key words frequency



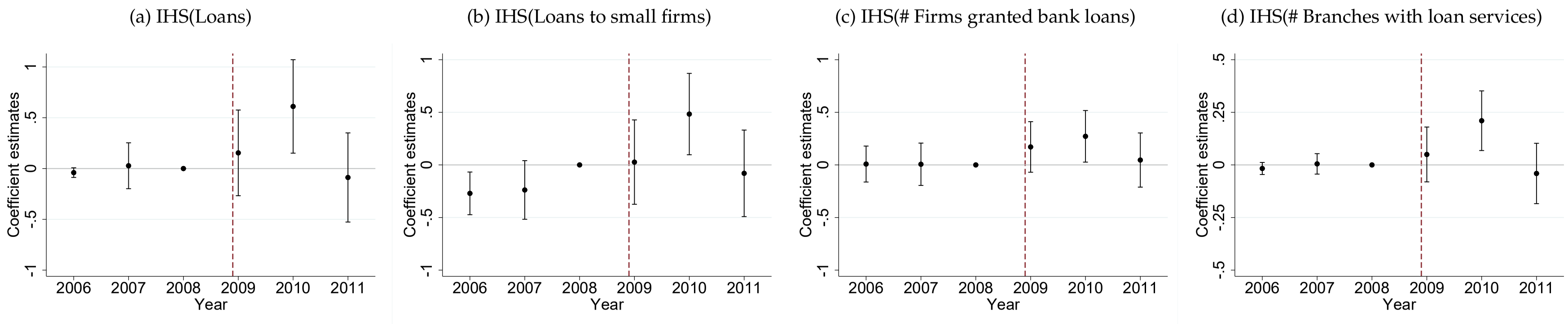
Machine learning-based topic score



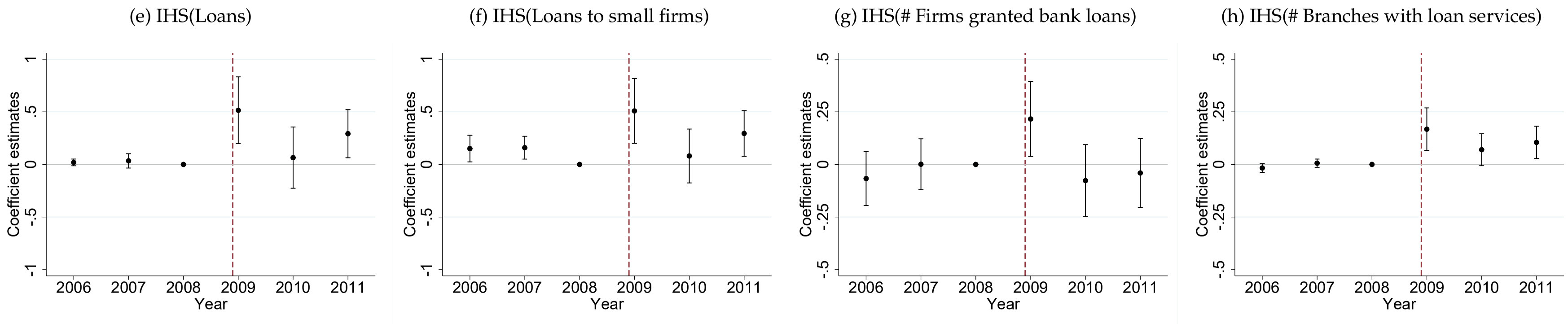
Notes: This figure shows the IV estimates on the dynamic effect of the reform on government policies across four policy areas. The unit of observation is county. The sample period is 2005-2018. The sample includes 97 counties. The estimation equations are event study variants of the specifications in Table 5 with the treatment instrumented by the randomly assigned rural survey teams in 1984. The year 2008, one year before the reform in 2009, is omitted as the reference year. Panels (a)-(e) measure policies using a simple key words frequency method. Panels (f)-(j) measure policies using supervised machine learning methods (Random Forest and Support Vector Machine). The detailed procedures for constructing these measures are described in Appendix C. To alleviate multiple hypothesis testing issues, panels (a) and (f) report estimates using a standardized index by summarizing the four policy measures following Kling, Liebman and Katz (2007). Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure A10: Dynamic effect on bank credit - IV estimates

Difference-in-differences estimates



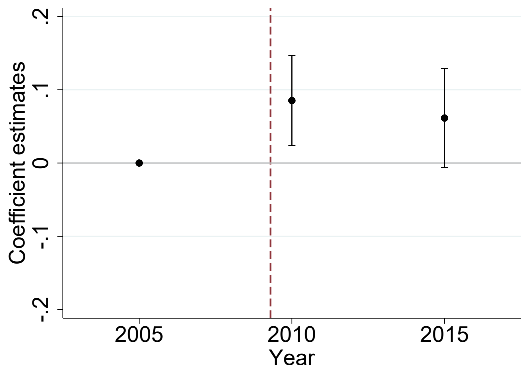
Difference-in-difference-in-differences estimates



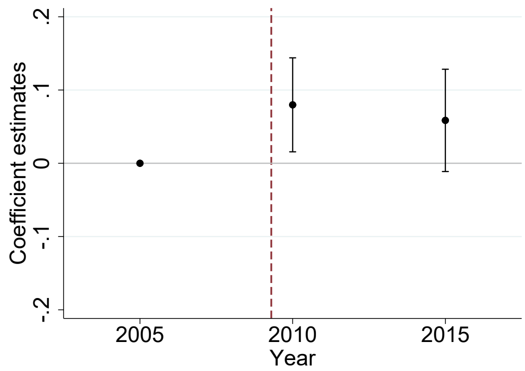
Notes: This figure shows IV estimates on the dynamic effect of the reform on bank credit. The unit of observation is county. The sample period is 2006-2011. The estimation equations are event study variants of the specifications in Table 6 with the treatment instrumented by the randomly assigned rural survey teams in 1984. The dependent variables are transformed by inverse hyperbolic sine (IHS) to reduce influences from the tails of the skewed outcome distributions (Bellemare and Wichman, 2020). Panels (a)-(d) conduct a conventional difference-in-differences estimation. Panels (e)-(h) further conduct a difference-in-difference-in-differences estimation utilizing differential control of banks by local governments across counties. Government control over banks is measured as the standardized share of pre-reform loans from City Commercial Banks (CCBs) in a county, whose controlling shareholders are local governments. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure A11: Dynamic effect on firm entry - IV estimates

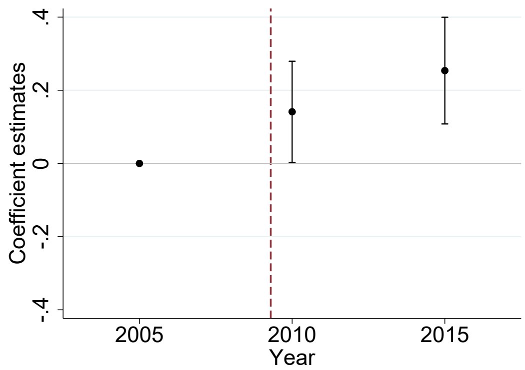
(a) IHS(# All firms)



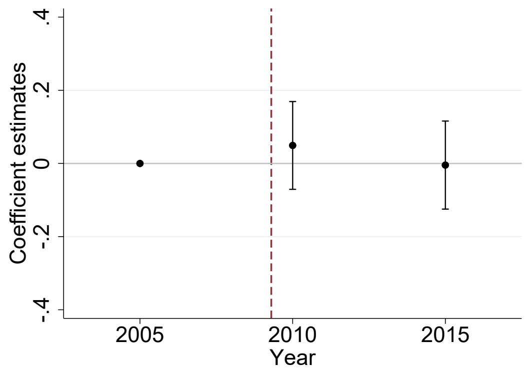
(b) IHS(# Private firms)



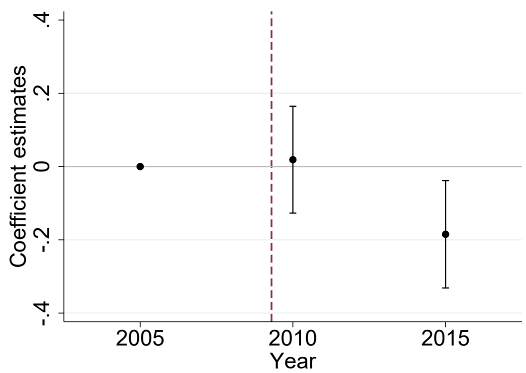
(c) IHS(# SOEs)



(d) IHS(# Foreign firms)

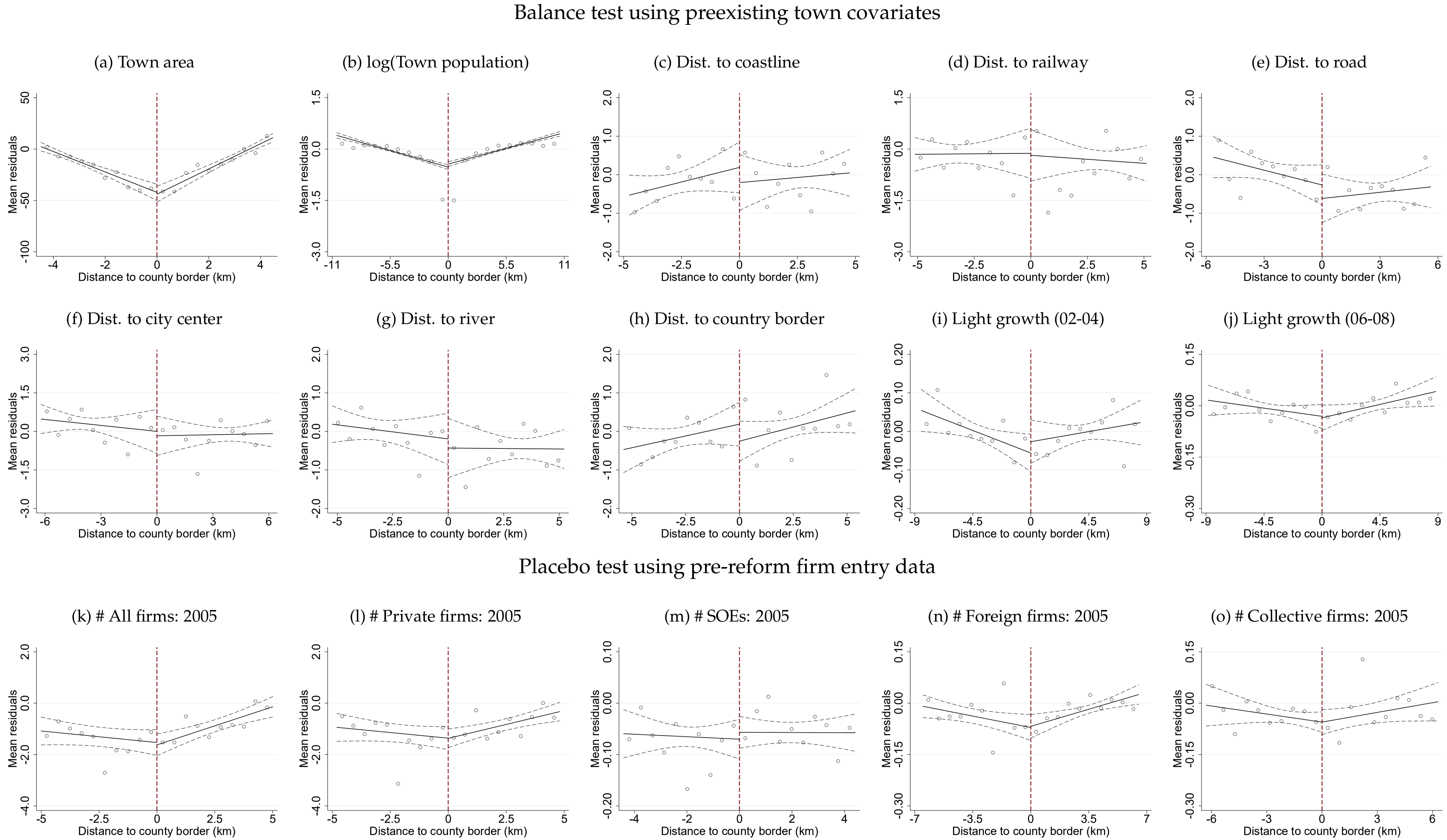


(e) IHS(# Collective firms)



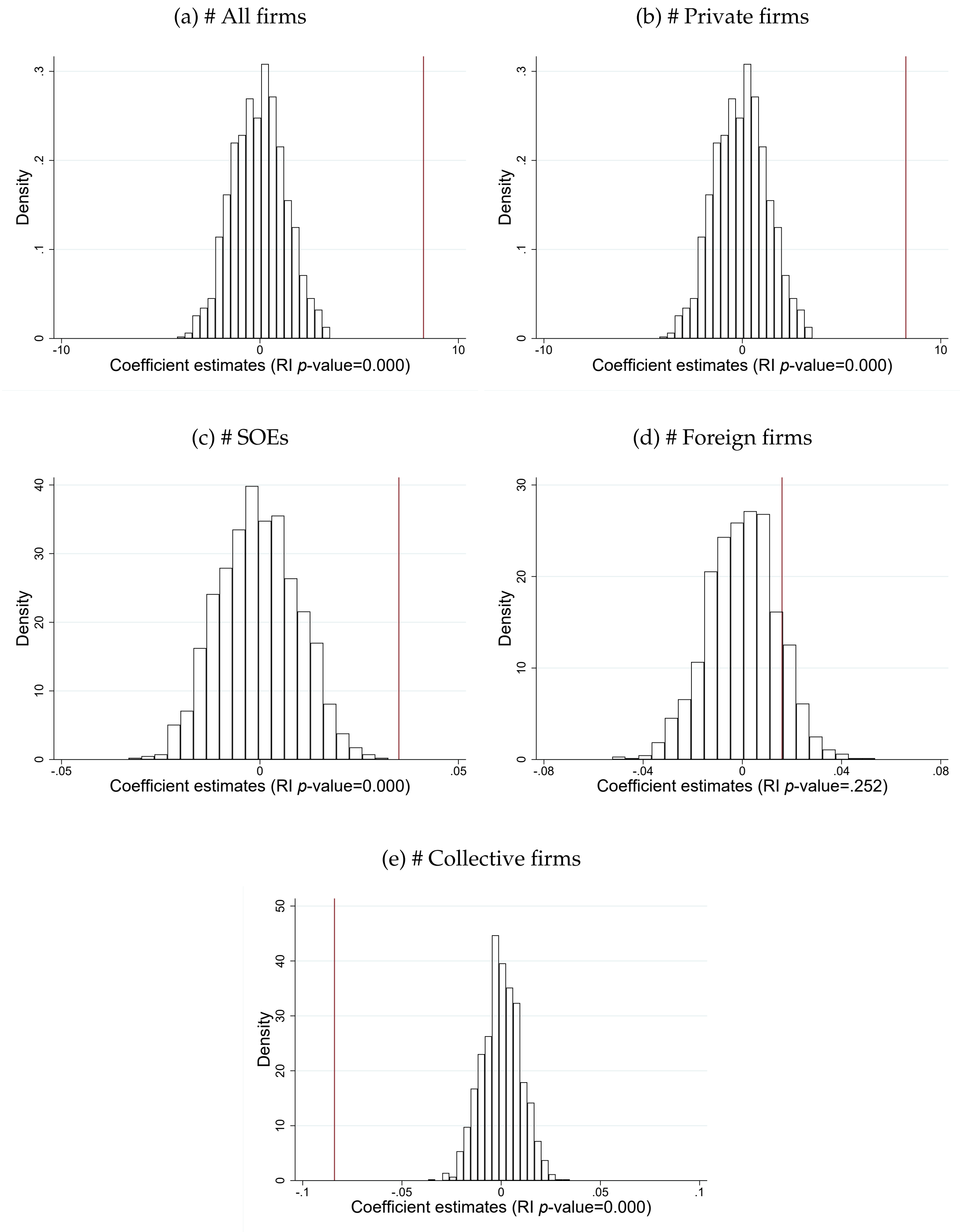
Notes: This figure shows the IV estimates on the dynamic effect of the reform on firm entry. The unit of observation is county. The sample includes the years 2005, 2010, and 2015. The estimation equations are event study variants of the specifications in Table 7 with the treatment instrumented by the randomly assigned rural survey teams in 1984. The dependent variables are transformed by inverse hyperbolic sine (IHS) to reduce influences from the tails of the skewed outcome distributions (Bellemare and Wichman, 2020). Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure A12: Effect on firm entry - RD balance check and placebo test



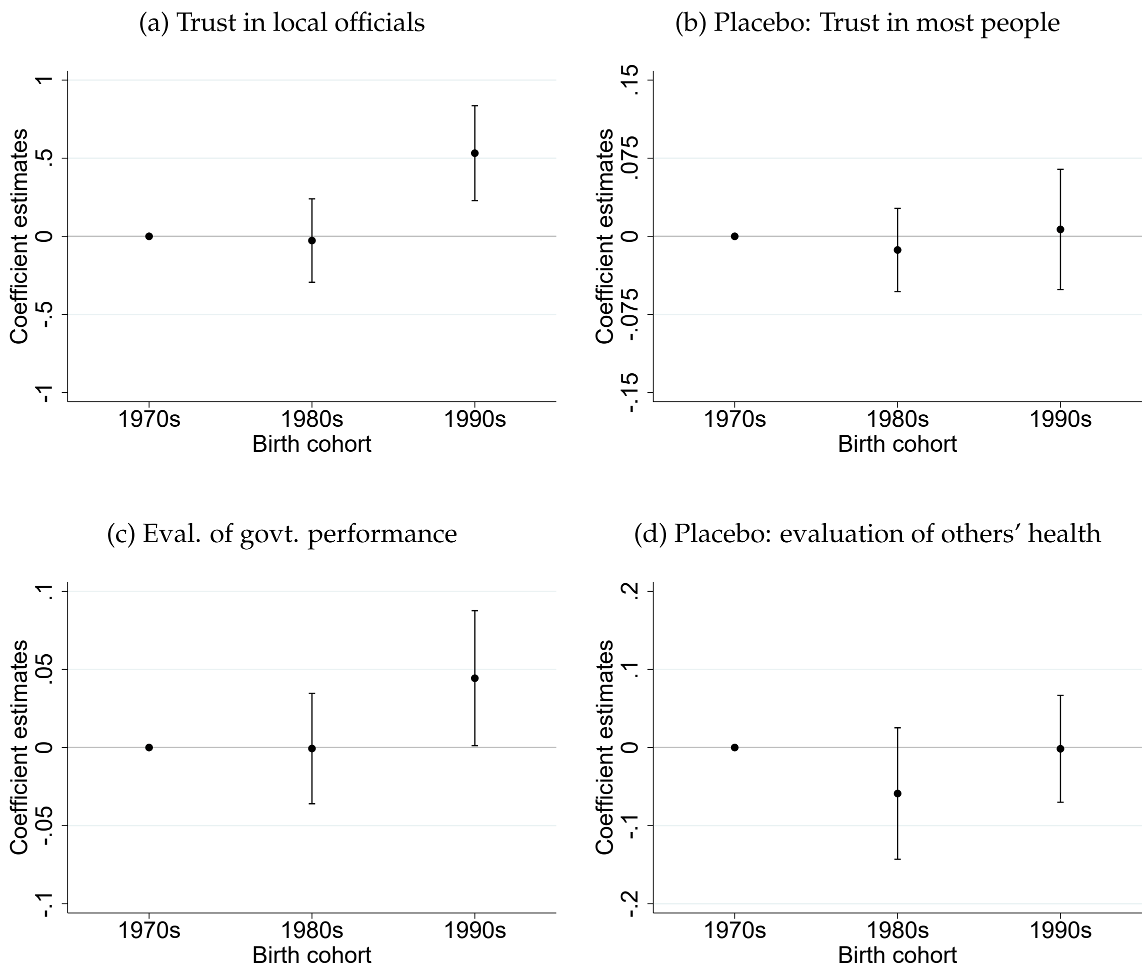
Notes: This figure conducts a balance check for the RD design using preexisting town covariates (panels a-j) and also a placebo test using pre-reform firm entry data (panels k-o). The unit of observation is town. The circles denote the mean of town covariates or firm entry within a distance bin, after partialling out county border fixed effects. The solid lines are fitted separately for each side of the county border. The dashes are the associated 90% confidence intervals constructed using standard errors clustered at the county border level. The bandwidths are the optimal bandwidths proposed by Calonico, Cattaneo and Titiunik (2014).

Figure A13: Effect on firm entry - Randomization inference for RD estimates



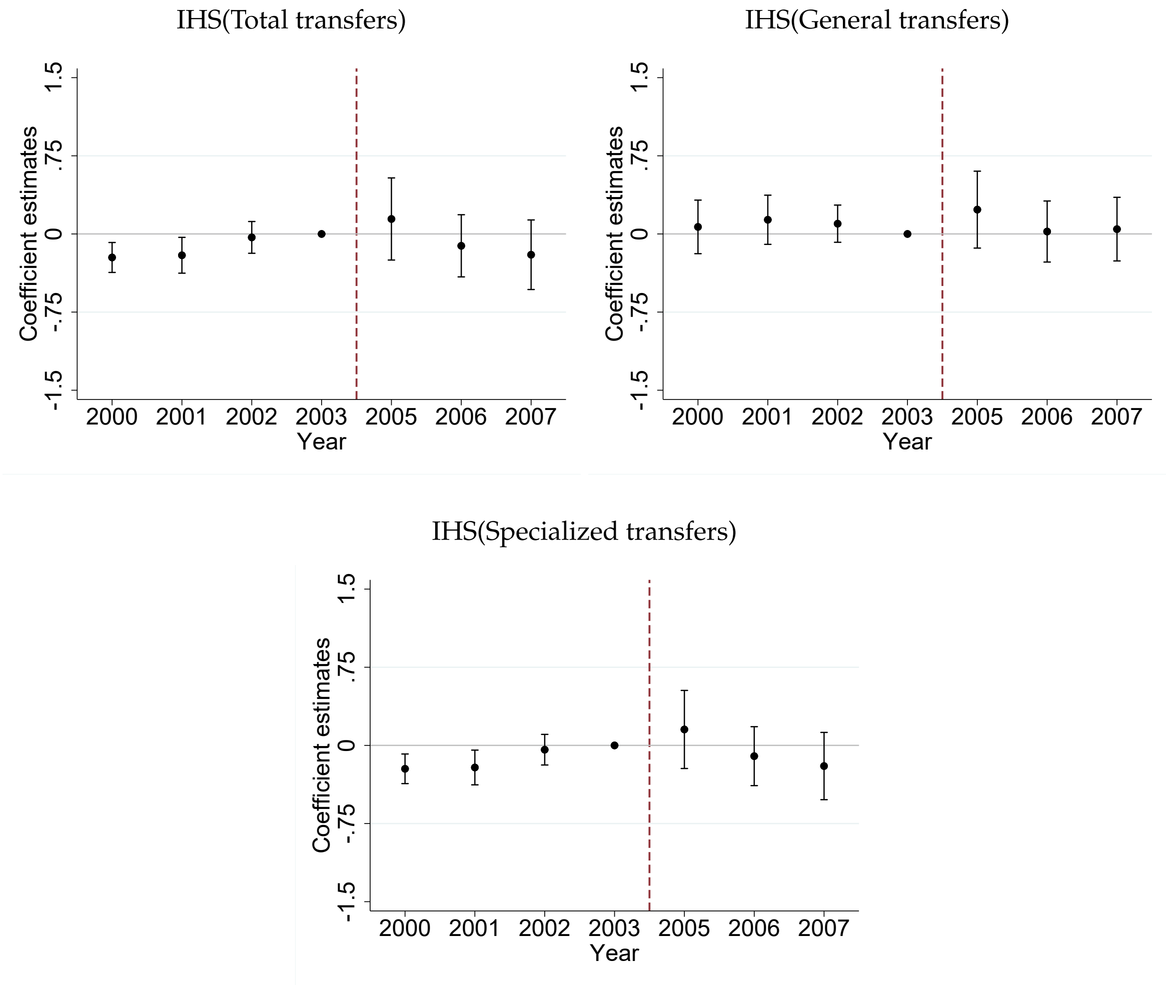
Notes: This figure adopts a randomization inference procedure with 2,000 permutations to show the robustness of the RD estimates following Ganong and Jäger (2018). Specifically, I keep only county borders across which there are no treatment variations, and then randomly create placebo treatment variation to each border. Based on these placebo borders, I then re-estimate the RD equation 13. This process is repeated for 2,000 times. The true estimates are denoted by the vertical lines and the randomization inference p-values are reported below the figures.

Figure A14: Effect on citizen attitudes - IV estimates



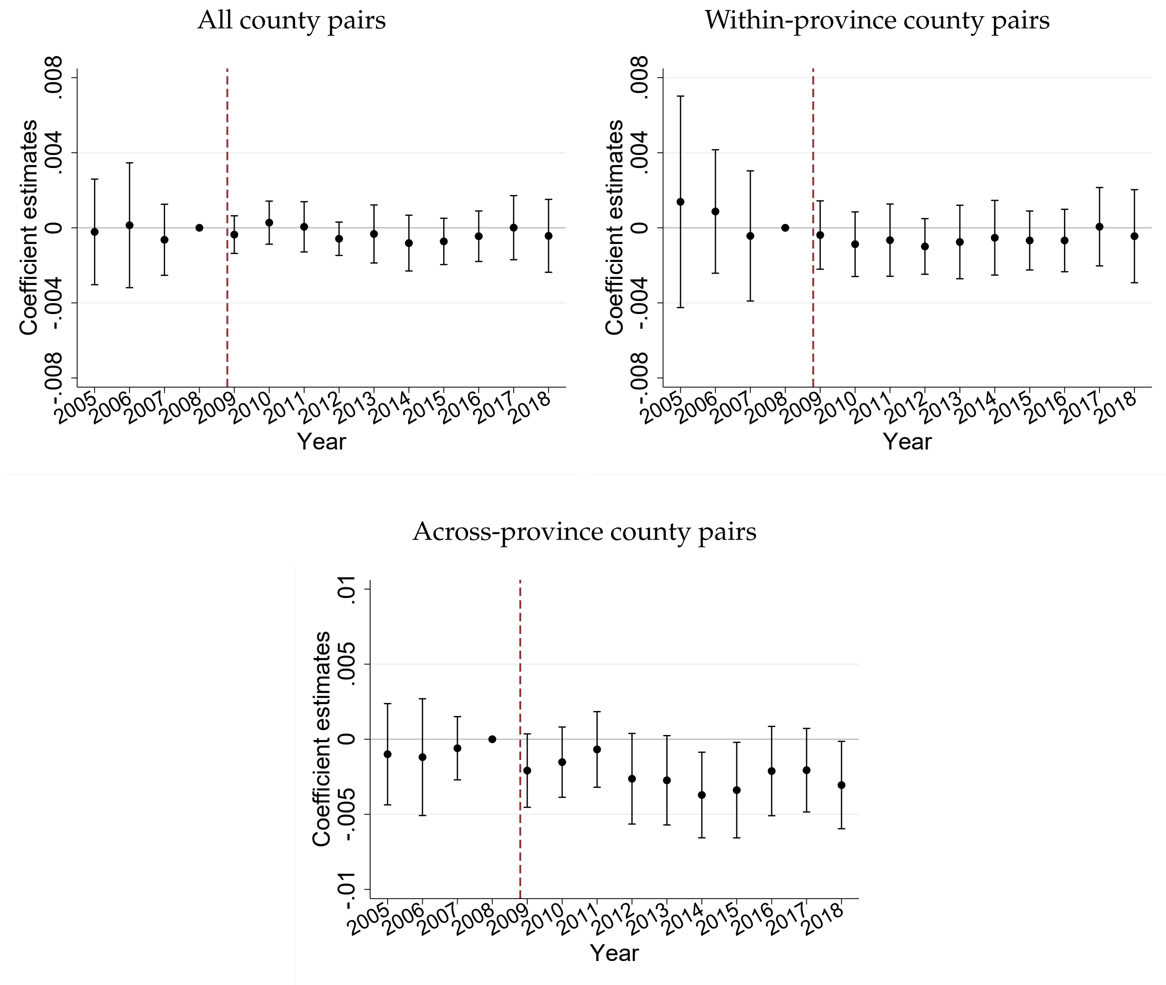
Notes: This figure shows the IV estimates on the effect of the reform on citizens' attitudes towards local governments, and is created by visualizing the results in Table A21. The unit of observation is citizen (three waves of survey data from the China Family Panel Studies, 2012, 2014 and 2016). The estimation equation is a cohort difference-in-differences specification (equation 14) that utilizes two sources of variation: (1) treatment counties versus control counties; (2) most affected cohorts versus less unaffected cohorts within the same county. The latter source of variation is built on insights in the psychology and political science literature that citizens' political attitudes are most permeable during teenage years and keep stable since one's 30s (Wolfinger and Rosenstone, 1980; Krosnick and Alwin, 1989). As such, the 1990s cohort are defined as the most affected group as they would be younger than 30 during the reform period, and thus, their political attitudes would be most permeable. The 1980s cohort are defined as the less affected group because some of them would be older than 30 during the reform period. The omitted group is those born in the 1970s, as they would be older than 30 in the reform period and thus be unaffected by the reform. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure A15: Dynamic effect on fiscal transfers around 2005



Notes: This table examines whether the launch of the survey teams would increase fiscal transfers from upper-level governments using an event study specification. The unit of observation is county. The sample is from 2000 to 2007. The dependent variables are transformed by inverse hyperbolic sine (IHS) to reduce influences from the tails of the skewed outcome distributions (Bellemare and Wichman, 2020). Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure A16: Dynamic effect on policy diffusion

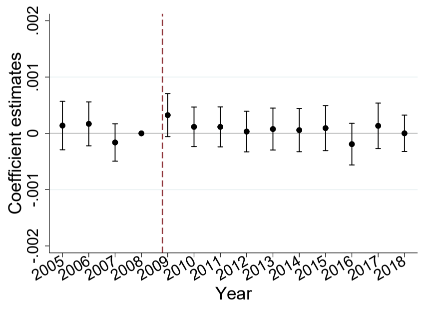


Notes: This figure shows the dynamic effect of the reform on policy diffusion across counties. The unit of observation is county pair. The dependent variable is the pairwise textual similarity of government work reports, which is calculated following Kelly et al. (2021) and described in Appendix C. Panel (a) considers all county pairs. Panel (b) considers county pairs within the same province. Panel (c) considers county pairs spanning different provinces. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are two-way clustered by both counties in a pair.

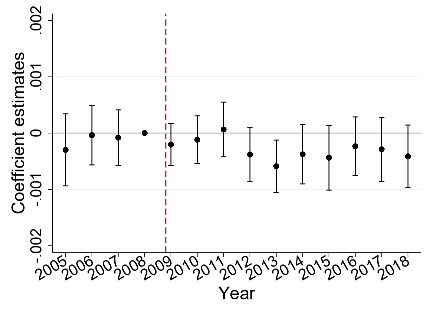
Figure A17: Effect on Government Policies - Social Welfare and Environmental Protection

Key words frequency

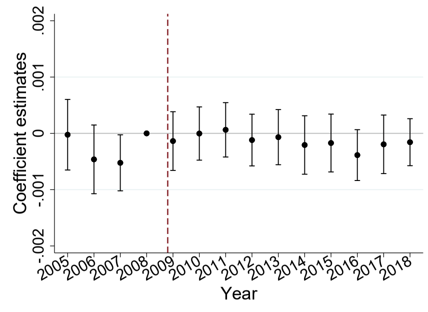
(a) Social welfare, OLS



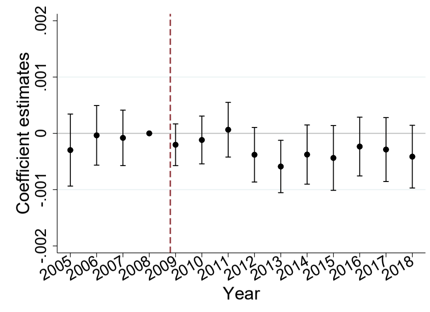
(b) Environmental protection, OLS



(c) Social welfare, IV

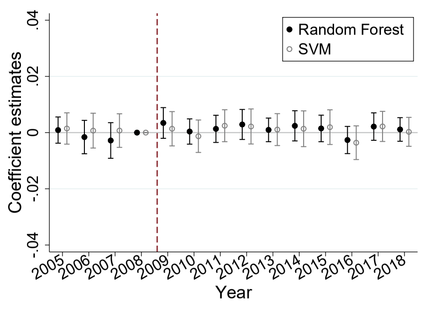


(d) Environmental protection, IV

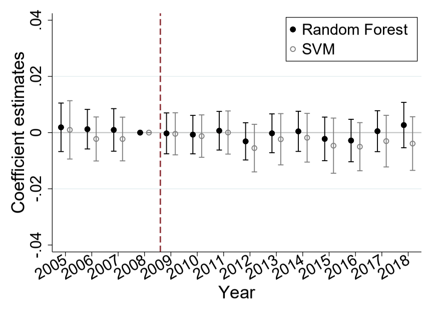


Machine learning-based topic score

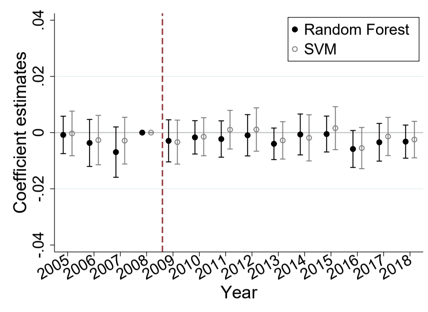
(e) Social welfare, OLS



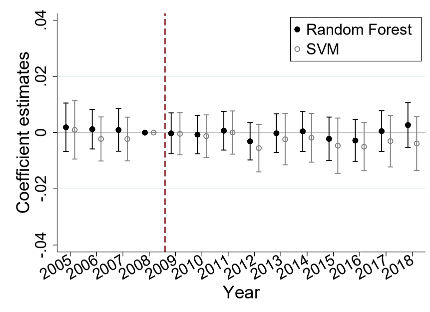
(f) Environmental protection, OLS



(g) Social welfare, IV



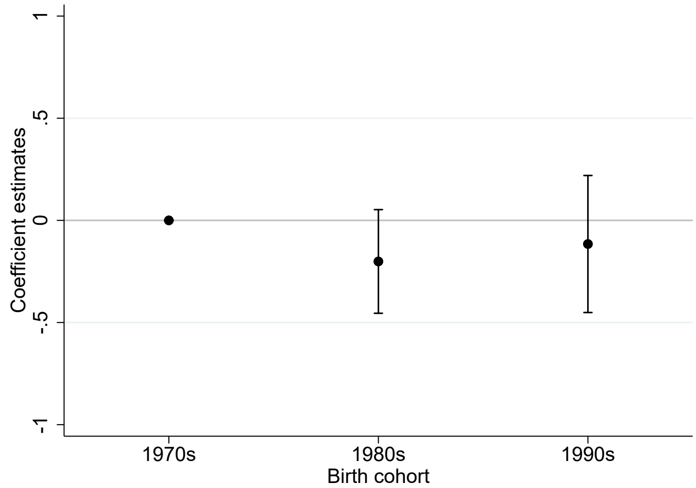
(h) Environmental protection, IV



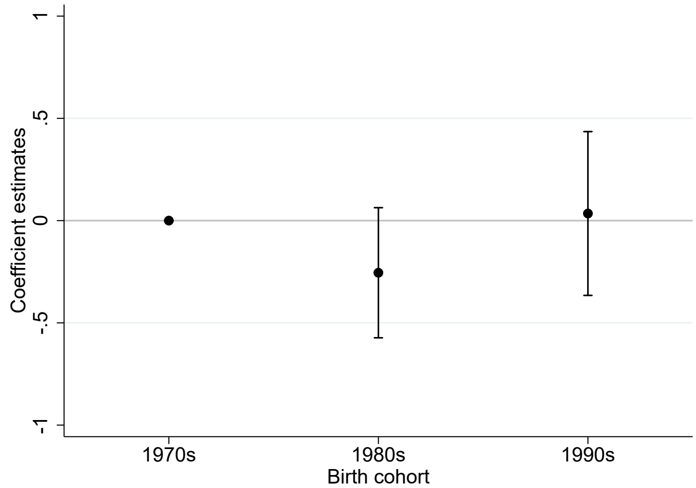
Notes: This figure shows the dynamic effect of the reform on government policies on social welfare and environmental protection. The unit of observation is county. The sample period is 2005-2018. The sample includes 97 counties. The policies are measured using either a simple key words frequency method or supervised machine learning methods (Random Forest and Support Vector Machine). The detailed procedures for constructing these measures are described in Appendix C. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

Figure A18: Effect on perceived corruption

(a) OLS estimates



(b) IV estimates



Notes: Notes: This figure shows the effect of the reform on citizens' perceived corruption about the government, and is created by visualizing the results in Table A28. The unit of observation is citizen (two waves of survey data from the China Family Panel Studies: 2014 and 2016). The estimation equation is a cohort difference-in-differences specification (equation 14) that utilizes two sources of variation: (1) treatment counties versus control counties; (2) most affected cohorts versus less unaffected cohorts within the same county. The latter source of variation is built on insights in the psychology and political science literature that citizens' political attitudes are most permeable during teenage years and keep stable since one's 30s (Wolfinger and Rosenstone, 1980; Krosnick and Alwin, 1989). As such, the 1990s cohort are defined as the most affected group as they would be younger than 30 during the reform period, and thus, their political attitudes would be most permeable. The 1980s cohort are defined as the less affected group because some of them would be older than 30 during the reform period. The omitted group is those born in the 1970s, as they would be older than 30 in the reform period and thus be unaffected by the reform. Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level.

E Additional tables

Table A1: Decomposing the effect of the reform

Dep. var.:	(1) OLS Reported GDP growth (%)	(2) OLS Light growth (%)	(3) IV Reported GDP growth (%)	(4) IV Light growth (%)
Treat × Post	-0.576***	0.254	-0.592***	0.162
	(0.161)	(0.294)	(0.203)	(0.355)
County FE	X	X	X	X
Year FE	X	X	X	X
Province FE × Post			X	X
Demographic controls × Post	X	X	X	X
Economic controls × Post	X	X	X	X
Geographic controls × Post	X	X	X	X
Cluster level	County	County	County	County
Observations	20,276	20,969	20,276	20,969
R-squared	0.361	0.364	0.100	0.020
Mean dep. var.	10.84	9.01	10.84	9.01
Effective *F*-statistic			2,278	2,287

Notes: This table shows the effect of the reform on reported GDP growth and light growth separately. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Columns (1)-(2) present OLS estimates. Columns (3)-(4) present IV estimates, with Treat instrumented by the randomly assigned rural survey teams in 1984. The effective *F*-statistics are constructed following Montiel Olea and Pflueger (2013). Standard errors clustered at the county level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Estimating spillover effect - robustness

	(1)	(2)	(3)	(4)
Dep. var.:	Reported GDP growth (%)			
Treat $\times$ Post	-0.610***	-0.559***	-0.581***	-0.576***
	(0.162)	(0.164)	(0.164)	(0.164)
# Treat neighbors within 50km $\times$ Post	-0.094			
	(0.091)			
# Treat neighbors within 100km $\times$ Post		0.027		
		(0.036)		
# Treat neighbors (GDP weighted) $\times$ Post			-0.011	
			(0.068)	
# Treat neighbors (population weighted) $\times$ Post				0.003
				(0.069)
Light growth (%)	0.017***	0.017***	0.017***	0.017***
	(0.005)	(0.005)	(0.005)	(0.005)
County FE	X	X	X	X
Year FE	X	X	X	X
County controls $\times$ Post	X	X	X	X
Neighbor number FE $\times$ Post	X	X	X	X
Cluster level	County	County	County	County
Mean dep. var.	10.84	10.84	10.84	10.84
Mean number of neighbors	2.40	11.45	5.91	5.91
Mean number of treat neighbors	0.88	4.47	1.97	1.97

Notes: The tables shows the robustness of the spillover effect of the reform on GDP growth manipulation. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. In columns (1) and (2), neighbors are defined as counties within a certain distance of a county, while in columns (3) and (4) neighbors are defined as counties sharing a common boundary segment with a county and are weighted by their sizes (GDP or population). Standard errors clustered at the county level are reported in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A3: Testing for monotonicity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep.var:	Treat × Post							
Group var.:	log Pop. (2010)		Share urban (%, 2010)		Share 15-64 (%, 2010)		Years of schooling (2010)	
	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)
$\text{Treat}^{1984} \times$ Post	0.747***	0.743***	0.768***	0.717***	0.762***	0.715***	0.748***	0.729***
	(0.024)	(0.021)	(0.022)	(0.025)	(0.023)	(0.025)	(0.024)	(0.024)
Observations	12,245	12,264	12,248	12,261	12,217	12,292	12,021	12,488
R-squared	0.905	0.900	0.900	0.881	0.897	0.881	0.892	0.885
Group var.:	Unem. rate (%, 2010)		Share 1st sectors (%, 2010)		Share 2nd sectors (%, 2010)		log GDP (2004)	
	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)
$\text{Treat}^{1984} \times$ Post	0.775***	0.703***	0.724***	0.751***	0.741***	0.737***	0.750***	0.748***
	(0.022)	(0.026)	(0.025)	(0.023)	(0.024)	(0.024)	(0.025)	(0.021)
Observations	12,049	12,460	12,250	12,259	12,231	12,278	11,251	13,258
R-squared	0.902	0.876	0.887	0.895	0.891	0.890	0.902	0.893
Group var.:	log GDP (2008)		GDP growth (%, 2002-2004)		GDP growth (%, 2006-2008)		Light growth (%, 2002-2004)	
	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)
$\text{Treat}^{1984} \times$ Post	0.755***	0.738***	0.725***	0.750***	0.725***	0.751***	0.736***	0.744***
	(0.023)	(0.022)	(0.025)	(0.023)	(0.025)	(0.024)	(0.024)	(0.024)
Observations	12,371	12,138	10,945	13,564	12,246	12,263	12,167	12,342
R-squared	0.905	0.893	0.888	0.890	0.883	0.894	0.887	0.892
Group var.:	Light growth (%, 2006-2008)		Dist. major roads (km, 2010)		Dist. major rail. (km, 2010)		County area (km2)	
	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)
$\text{Treat}^{1984} \times$ Post	0.731***	0.747***	0.765***	0.716***	0.778***	0.711***	0.779***	0.701***
	(0.024)	(0.024)	(0.023)	(0.025)	(0.023)	(0.025)	(0.023)	(0.025)
Observations	12,210	12,299	11,997	12,512	12,068	12,441	12,180	12,329
R-squared	0.886	0.892	0.898	0.885	0.897	0.881	0.903	0.881
Group var.:	Precipitation (inches, 2004)		Temperature (degrees, 2004)		Precipitation (inches, 2008)		Temperature (degrees, 2008)	
	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)
$\text{Treat}^{1984} \times$ Post	0.739***	0.739***	0.730***	0.743***	0.726***	0.752***	0.729***	0.746***
	(0.024)	(0.024)	(0.025)	(0.023)	(0.025)	(0.023)	(0.025)	(0.023)
Observations	12,301	12,208	12,273	12,236	12,301	12,208	12,273	12,236
R-squared	0.891	0.893	0.887	0.891	0.887	0.898	0.886	0.891
Group var.:	Dist. to major rivers (km)		Dist. to country border (km)		Dist. to coastline (km)		Dist. to city center (km)	
	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)	<p(50)	>=p(50)
$\text{Treat}^{1984} \times$ Post	0.776***	0.702***	0.706***	0.778***	0.740***	0.733***	0.753***	0.734***
	(0.022)	(0.026)	(0.025)	(0.022)	(0.024)	(0.025)	(0.023)	(0.024)
Observations	12,249	12,260	12,264	12,245	12,180	12,329	12,213	12,296
R-squared	0.903	0.875	0.879	0.899	0.893	0.883	0.891	0.888

Notes: This table provides evidence to support the monotonicity assumption required for IV estimation in the presence of heterogeneous treatment effects. It does so by reporting the first-stage results for subsamples divided by the medians of baseline county covariates. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Treat^{1984} is a dummy variable indicating counties with a randomly assigned rural survey team in 1984. Post is a dummy variable indicating years after the reform in 2009. The estimation equation is equation (9). County fixed effects, year fixed effects, province fixed effects interacted the post-reform dummy, and a control for light growth are included but not shown due to space limitation. Standard errors clustered at the county level are reported in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A4: Reweighting OLS to match IV

	(1)	(2)	(3)	(4)	(5)
Dep. var.:	Reported GDP growth (%)				
Treat × Post	-0.576***	-0.565***	-0.568***	-0.567***	-0.535***
	(0.161)	(0.161)	(0.161)	(0.162)	(0.162)
Light growth (%)	0.017***	0.016***	0.016***	0.016***	0.016***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
County FE	X	X	X	X	X
Year FE	X	X	X	X	X
County controls × Post	X	X	X	X	X
Reweighting		X	X	X	X
Cluster level	County	County	County	County	County
Observations	20,273	20,273	20,273	20,273	20,273
R-squared	0.362	0.361	0.361	0.361	0.360
Number of subgroups	-	10	20	30	2x2x2x2x2
Mean dep. var.	10.84	10.84	10.84	10.84	10.84

Notes: This table adopts a reweighting method to adjust the OLS estimates to match the sample of compliers. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Column (1) reproduces the baseline OLS estimates for comparison. Columns (2)-(4) divides the sample into 10-30 groups of equal size, based on quantiles of the first principal component of baseline county covariates, and then reweight the raw OLS estimation using the complier share in each group as weights. Column (5) divides the sample into 32 groups using the medians of the first five principal components of baseline county covariates, and then reweights the OLS using the complier share in each group as weights. Standard errors clustered at the county level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Effect in a trimmed sample without personnel changes

	(1)	(2)	(3)
Dep. var.:		Reported GDP growth (%)	
Sample:	Baseline	Terms straddling 2009	Terms covering 2007-2011
Treat × Post	-0.576***	-0.617***	-0.647***
	(0.161)	(0.207)	(0.207)
Light growth rate (%)	0.017***	0.017***	0.016***
	(0.005)	(0.005)	(0.005)
County FE	X	X	X
Year FE	X	X	X
County controls × Post	X	X	X
Cluster level	County	County	County
Observations	20,273	16,722	16,384
R-squared	0.362	0.370	0.367
Mean dep. var.	10.84	10.98	10.94

Notes: This table examines whether the reduction in manipulation stemmed from personnel changes among local officials, by utilizing a trimmed sample in which there were no personnel changes in the treatment counties. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Standard errors clustered at the county level are reported in parentheses. * p <0.1, ** p <0.05, *** p <0.01.

Table A6: Effects on leader traits and turnovers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dep. var:	Years in office	Term length	Age	1(Local)	Schooling	Connection	1(Turnover)
Panel A: Magistrates							
Treat × Post	0.046	0.104	0.048	-0.005	-0.147	0.052*	0.001
	(0.160)	(0.071)	(0.230)	(0.017)	(0.112)	(0.028)	(0.007)
Observations	22,628	22,628	9,876	7,626	7,038	8,211	22,628
R-squared	0.881	0.975	0.866	0.873	0.858	0.504	0.173
Mean dep. var.	2.49	4.53	45.13	0.19	16.02	0.24	0.06
Panel B: Party secretaries							
Treat × Post	0.113	0.029	-0.075	-0.009	-0.157	-0.029	-0.001
	(0.163)	(0.071)	(0.294)	(0.016)	(0.104)	(0.028)	(0.006)
Observations	22,695	22,695	10,674	8,243	7,994	9,027	22,695
R-squared	0.879	0.976	0.769	0.857	0.829	0.475	0.151
Mean dep. var.	2.73	4.86	47.68	0.16	16.40	0.24	0.05
County FE	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X
Cluster level	County	County	County	County	County	County	County

Notes: This table test the effects of the reform on personnel turnovers and personnel traits. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Standard errors clustered at the county level are reported in parentheses. * p <0.1, ** p <0.05, *** p <0.01.

Table A7: Estimating promotion incentives

	(1)	(2)	(3)	(4)
Dep. var.:	1(Promotion)			
Official type:	Party secretary		Magistrate	
Model:	LPM	Probit	LPM	Probit
Start age	-0.033***	-0.234***	-0.017**	-0.176***
	(0.009)	(0.056)	(0.007)	(0.065)
Connection	0.022**	0.142***	0.019**	0.139**
	(0.009)	(0.053)	(0.008)	(0.066)
Education	0.008	0.035	0.017**	0.152*
	(0.009)	(0.062)	(0.007)	(0.087)
Start age × Connection	0.012	0.107**	0.005	0.056
	(0.009)	(0.049)	(0.007)	(0.055)
Start age × Education	-0.002	-0.012	0.001	0.018
	(0.010)	(0.059)	(0.008)	(0.077)
Connection × Education	0.020**	0.105**	0.017**	0.117*
	(0.009)	(0.053)	(0.008)	(0.071)
Observations	1,093	1,093	1,018	1,018

Notes: This table estimates the promotion incentives of county leaders. The unit of observation is local leader, either the party secretary or the magistrate in a county. The sample period is 2005-2018. The promotion incentives denote local leaders' ex ante likelihood of promotion based on their start ages, years of schooling, and political connections with upper-level leaders following Avis, Ferraz and Finan (2018) and Wang, Zhang and Zhou (2020). Odd columns adopt linear probability models and even columns adopt Probit models. Note that other performance variables, such as GDP growth, are intentionally excluded from this regression, so the estimated probabilities capture the ex ante likelihood of promotion (Wang, Zhang and Zhou, 2020). Robust standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: Testing the promotional discipline effect

	(1)	(2)	(3)
Dep. var.:		Reported GDP growth (%)	
Treat × Post	-0.576***	-0.658***	-0.562**
	(0.161)	(0.243)	(0.269)
Light growth rate (%)	0.017***	0.019***	0.011
	(0.005)	(0.007)	(0.008)
Treat × Post × 1(Age>52)		0.253	
		(1.164)	
1(Age>52)		-0.931	
		(0.687)	
Treat × 1(Age>52)		-0.081	
		(1.129)	
Post × 1(Age>52)		0.535	
		(0.709)	
Treat × Post × Promotion incentive			0.127
			(0.217)
Promotion incentive			-0.112
			(0.138)
Treat × Promotion incentive			-0.027
			(0.210)
Post × Promotion incentive			0.042
			(0.146)
County FE	X	X	X
Year FE	X	X	X
Cluster level	County	County	County
Observations	20,273	17,980	11,989
R-squared	0.362	0.355	0.370
Mean dep. var.	10.84	10.67	10.77

Notes: This table tests the promotional discipline effect, by checking whether the reform effect is larger for local officials with greater promotion incentives. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Promotion incentive denotes local officials' ex ante likelihood of promotion estimated based on their start ages, years of schooling, and political connections with upper-level leaders (Avis, Ferraz and Finan, 2018; Wang, Zhang and Zhou, 2020), and are estimated separately for the party secretary and the magistrate. Column (1) reproduces the baseline estimate. In columns (2)-(3), I adopt a slightly different specification to allow the two leaders in a county (the party secretary and the magistrate) to separately affect GDP growth manipulation. This means that for each county-year, I generate two parallel observations that are identical except for one distinction: one includes only the party secretary, while the other includes only the magistrate. Standard errors clustered at the county level are reported in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A9: Testing the soft information channel

	(1)	(2)	(3)
Dep. var.:		Reported GDP growth (%)	
Treat × Post	-0.576***	-0.575***	-0.549*
	(0.161)	(0.163)	(0.287)
Light growth (%)	0.017***	0.017***	0.015**
	(0.005)	(0.005)	(0.007)
Treat × Post × Distance to upper-level govt.		0.009	
		(0.175)	
Treat × Post × Connection			-0.057
			(0.069)
Connection			0.058
			(0.046)
Treat × Connection			-0.073
			(0.064)
Post × Connection			-0.030
			(0.050)
County FE	X	X	X
Year FE	X	X	X
County controls × Post	X	X	X
Cluster level	County	County	County
Observations	20,273	20,273	10,802
R-squared	0.362	0.362	0.365
Mean of dep. var.	10.84	10.84	10.64

Notes: This table tests whether the reduction in manipulation was driven by soft information provided by the survey teams. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Distance to upper-level govt. denotes the shortest distance from a county to its overseeing upper-level government. Connection is a dummy variable that equals 1 if the party secretary or magistrate in a county shares the same hometown or educational background with upper-level leaders, and 0 otherwise. Standard errors clustered at the county level are reported in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A10: Effect on statistical capacity

	(1)	(2)
Dep. var.:	1(Award for outstanding performance in economic census)	
Treat × Post	0.011	-0.005
	(0.017)	(0.020)
County FE	X	X
Year FE	X	X
County controls × Post		X
Cluster level	County	County
Observations	7,116	5,992
R-squared	0.391	0.420
Mean of dep. var.	0.13	0.13

Notes: This table shows the effect of the reform on local statistical capacity. The unit of observation is county. The sample period is 2004, 2008, 2013, and 2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. The dependent variable is a dummy denoting whether a county won an award for outstanding performance in conducting economic census. Standard errors clustered at the county level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Controlling for concurrent reforms

	(1)	(2)	(3)	(4)
Dep. var.:	Reported GDP growth (%)			
Treat × Post	-0.593***	-0.579***	-0.576***	-0.596***
	(0.162)	(0.161)	(0.161)	(0.162)
Light growth (%)	0.016***	0.017***	0.017***	0.016***
	(0.005)	(0.005)	(0.005)	(0.005)
Fiscal PMC	X			X
Full PMC		X		X
Anticorruption inspection			X	X
County FE	X	X	X	X
Year FE	X	X	X	X
County controls × Post	X	X	X	X
Cluster level	County	County	County	County
Observations	20,273	20,273	20,273	20,273
R-squared	0.362	0.362	0.362	0.363
Mean of dep. var.	10.84	10.84	10.84	10.84

Notes: This table shows the robustness of te baseline results after controlling for several concurrent reforms that may also strengthen the monitoring of local officials. The unit of observation is county. The sample period is 2005-2018. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Fiscal PMC is a dummy for the fiscal province-managing-county (PMC) reform. Full PMC is a dummy for the full province-managing-county (PMC) reform. Standard errors clustered at the county level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Keywords in each policy

Policy	Chinese keywords	English translation
business attraction	招商引资	attract businesses
	外商直接投资	foreign direct investment
	外资	foreign capital
	对外开放	open up
infrastructure	基础设施	infracstrcture
	基建	shorthand term for infractrcture
	工程建设	project construction
	建设项目	construction project
market reform	改革	reform
	非公有制	non-public ownership
	民营企业	private firms
	私营企业	private firms [variant]
	民企	shorthand term for private firms
	私企	shorthand term for private firms [variant]
policy experimentation	试点	experimental places
	试验区	experimental zones
social welfare	社会保险	social insurance
	社保	shorthand term for social insurance
	养老保险	endowment insurance
	医疗保险	medical insurance
	养老金	pension
	社会保障	social security
Environmental protection	环境保护	environmental protection
	环保	shorthand term for environmental protection
	污染治理	pollution control
	减排	emission reduction

Notes: This table lists the keywords in each policy, which is used to create keywords frequency measures of local officials' emphasis on each policy. The first four policies are used in Section 5.1. The last two policies are examined in Section 5.5.

Table A13: Effect on government policies - IV estimates

Dep. var.:	(1) Standardized index	(2) Business attraction	(3) Infrastructure investment	(4) Market reform	(5) Policy experimentation
Panel A: Key words frequency					
Treat × Post	0.664***	0.001***	-0.000	0.001***	0.000
	(0.201)	(0.000)	(0.000)	(0.000)	(0.000)
R-squared	0.114	0.139	0.105	0.096	0.083
Mean dep. var.	0	0.002	0.002	0.004	0.001
Panel B: Topic score predicted by Random Forest					
Treat × Post	0.539***	0.007***	-0.005	0.011*	0.007**
	(0.188)	(0.002)	(0.005)	(0.006)	(0.003)
R-squared	0.063	0.091	0.047	0.047	0.059
Mean dep. var.	0	0.013	0.040	0.068	0.016
Panel C: Topic score predicted by Support Vector Machine					
Treat × Post	0.566***	0.007**	-0.004	0.011**	0.007**
	(0.195)	(0.003)	(0.005)	(0.005)	(0.003)
R-squared	0.088	0.118	0.063	0.055	0.084
Mean dep. var.	0	0.014	0.051	0.068	0.019
County FE	X	X	X	X	X
Year FE	X	X	X	X	X
Province FE × Post	X	X	X	X	X
County controls × Post	X	X	X	X	X
Cluster level	County	County	County	County	County
Observations	883	883	883	883	883
Effective *F*-statistic	74.06	74.06	74.06	74.06	74.06

Notes: This table shows the IV estimates on the effect of the reform on government policies across four policy areas. The unit of observation is county. The sample period is 2005-2018. The sample includes 97 counties. Treat is a dummy variable indicating counties with the survey teams deployed in 2005, and is instrumented by the randomly assigned rural survey teams in 1984. Post is a dummy variable indicating years after the reform in 2009. Panel A measures policies using a simple key words frequency method. Panel B and C measure policies using supervised machine learning methods (Random Forest and Support Vector Machine). The detailed procedures for constructing these measures are described in Appendix C. To alleviate multiple hypothesis testing issues, column (1) reports estimates using an s-tandardized index by summarizing the four policy measures following Kling, Liebman and Katz (2007). The effective F-statistics are constructed following Montiel Olea and Pflueger (2013). Standard errors clustered at the county level are reported in parentheses. * p <0.1, ** p <0.05, *** p <0.01.

Table A14: Effect on bank credit - IV estimates

Dep. var. is IHS of:	(1) Total amount of loans	(2) Loans to small firms	(3) # Firms granted loans	(4) # Branches granting loans
Panel A: Difference-in-differences				
Treat × Post	0.190	0.270*	0.138	0.090*
	(0.158)	(0.150)	(0.095)	(0.046)
Effective F-statistic	2,058	2,058	2,058	2,058
Panel B: Difference-in-difference-in-differences				
Treat × Government control × Post	0.273**	0.191*	0.055	0.118***
	(0.118)	(0.103)	(0.065)	(0.037)
Treat × Post	0.223	0.309**	0.160*	0.073
	(0.155)	(0.152)	(0.096)	(0.046)
Government control × Post	0.191**	0.172**	0.141***	0.059**
	(0.086)	(0.076)	(0.046)	(0.028)
Effective F-statistic	1,022	1,022	1,022	1,022
County FE	X	X	X	X
Year FE	X	X	X	X
Province FE × Post	X	X	X	X
County controls × Post	X	X	X	X
Observations	8,922	8,922	8,922	8,922

Notes: This table shows the IV estimates on the effect of the reform on bank credit. The unit of observation is county. The sample period is 2006-2011. Treat is a dummy variable indicating counties with the survey teams deployed in 2005, and is instrumented by the randomly assigned rural survey teams in 1984. Post is a dummy variable indicating years after the reform in 2009. Government control denotes the standardized share of pre-reform loans from City Commercial Banks (CCBs) in a county, whose controlling shareholders are local governments. The dependent variables are highly skewed and are thus transformed by inverse hyperbolic sine (IHS) to reduce influences from the tails (Bellemare and Wichman, 2020). Panel A adopts a difference-in-differences estimation as usual. Panel B adopts a difference-in-difference-in-differences estimation utilizing differential control of banks by local governments across counties. The estimation equation is based on equation (12). The effective F-statistics are constructed following Montiel Olea and Pflueger (2013). Standard errors clustered at the county level are reported in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A15: Effect on bank credit - untransformed variables

Dep. var.:	(1) Total amount of loans	(2) Loans to small firms	(3) # Firms granted loans	(4) # Branches granting loans
Panel A: Difference-in-differences				
Treat × Post	24,960.061**	8,911.357**	31.685	1.012*
	(10,783.378)	(3,741.146)	(31.821)	(0.558)
R-squared	0.881	0.765	0.611	0.859
Panel B: Difference-in-difference-in-differences				
Treat × Government control × Post	84,666.050*	19,681.718	-608.650	1.311*
	(50,782.638)	(13,658.901)	(540.779)	(0.691)
Treat × Post	129,932.700***	25,907.130***	-376.102	-0.248
	(33,496.942)	(9,745.098)	(323.677)	(0.585)
Government control × Post	67,981.316***	18,991.578***	683.584	0.655*
	(23,239.581)	(6,560.406)	(535.012)	(0.358)
R-squared	0.877	0.747	0.188	0.891
County controls × Post	X	X	X	X
County FE	X	X	X	X
Year FE	X	X	X	X
Observations	8,922	8,922	8,922	8,922
Mean dep. var.	368,082	78,087	525	41.11

Notes: This table shows the effect of the reform on bank credit, without inverse hyperbolic sine transformation of the outcomes to alleviate concerns raised by Chen and Roth (2023). The unit of observation is county. The sample period is 2006-2011. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Government control denotes the standardized share of pre-reform loans from City Commercial Banks (CCBs) in a county, whose controlling shareholders are local governments. Panel A adopts a difference-in-differences estimation as usual. Panel B adopts a difference-in-difference-in-differences estimation utilizing differential control of banks by local governments across counties. The estimation equation is equation (12). Standard errors clustered at the county level are reported in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A16: Effect on firm entry - untransformed variables

	(1)	(2)	(3)	(4)	(5)
Dep. var.:			# Firm registrations		
Firm type:	All	Private	SOEs	Foreign	Collective
Treat × Post	23.635**	23.173**	0.743***	0.144	-0.426
	(10.976)	(10.933)	(0.264)	(0.213)	(0.305)
County FE	X	X	X	X	X
Year FE	X	X	X	X	X
County controls × Post	X	X	X	X	X
Cluster level	County	County	County	County	County
Observations	4,494	4,494	4,494	4,494	4,494
R-squared	0.834	0.833	0.496	0.811	0.505
Mean dep. var.	305.6	298.1	3.358	2.305	1.831

Notes: This table shows the effect of the reform on firm entry, without inverse hyperbolic sine transformation of the outcomes to alleviate concerns raised by Chen and Roth (2023). The unit of observation is county. The sample includes the years 2005, 2010, and 2015. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. Standard errors clustered at the county level are reported in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A17: Effect on firm entry - IV estimates

	(1)	(2)	(3)	(4)	(5)
Dep. var.:			IHS(# Firm registrations)		
Firm type:	All	Private	SOEs	Foreign	Collective
Treat × Post	0.073**	0.069**	0.197***	0.022	-0.083
	(0.034)	(0.035)	(0.073)	(0.064)	(0.078)
County FE	X	X	X	X	X
Year FE	X	X	X	X	X
Province FE × Post	X	X	X	X	X
County controls × Post	X	X	X	X	X
Cluster level	County	County	County	County	County
Observations	4,494	4,494	4,494	4,494	4,494
R-squared	0.057	0.068	0.019	0.017	0.033
Effective *F*-statistic	2,260	2,260	2,260	2,260	2,260

Notes: This table shows the IV estimates on the effect of the reform on firm entry. The unit of observation is county. The sample includes the years 2005, 2010, and 2015. Treat is a dummy variable indicating counties with the survey teams deployed in 2005, and is instrumented by the randomly assigned rural survey teams in 1984. Post is a dummy variable indicating years after the reform in 2009. The effective *F*-statistics are constructed following Montiel Olea and Pflueger (2013). Standard errors clustered at the county level are reported in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table A18: Effect on firm entry - RD robustness

	(1)	(2)	(3)	(4)	(5)
Dep. var.:			# Firm registrations		
Firm type:	All	Private	SOEs	Foreign	Collective
Panel A: IK optimal bandwidth					
Treat	7.515*	7.953**	0.045	0.055*	-0.049
	(3.954)	(3.830)	(0.027)	(0.033)	(0.040)
RD kernel	Uniform	Uniform	Uniform	Uniform	Uniform
RD polynomial	1	1	1	1	1
Rd bandwidth	IK	IK	IK	IK	IK
Observations	8,716	8,954	10,480	11,092	6,938
R-squared	0.354	0.349	0.205	0.427	0.254
Mean dep. var.	18.49	18.24	0.11	0.08	0.05
Bandwidth	3.97	4.07	4.74	5.05	3.30
Panel B: Quadratic RD polynomial					
Treat	8.448*	8.225*	0.084**	0.146	-0.090**
	(4.446)	(4.432)	(0.034)	(0.123)	(0.038)
RD kernel	Uniform	Uniform	Uniform	Uniform	Uniform
RD polynomial	2	2	2	2	2
Rd bandwidth	CCT	CCT	CCT	CCT	CCT
Observations	13,800	13,686	15,096	16,966	16,698
R-squared	0.266	0.266	0.164	0.501	0.212
Mean dep. var.	20.15	19.88	0.14	0.11	0.09
Bandwidth	6.71	6.62	7.87	10.06	9.71
Panel C: Triangular RD polynomial					
Treat	8.897**	8.877**	0.042*	0.059**	-0.073**
	(3.660)	(3.661)	(0.024)	(0.029)	(0.031)
RD kernel	Triangular	Triangular	Triangular	Triangular	Triangular
RD polynomial	1	1	1	1	1
Rd bandwidth	CCT	CCT	CCT	CCT	CCT
Observations	11,126	10,936	12,548	14,820	12,408
R-squared	0.341	0.341	0.218	0.375	0.214
Mean dep. var.	18.78	18.26	0.12	0.09	0.08
Bandwidth	5.07	4.98	5.83	7.60	5.74
County border FE	X	X	X	X	X
Cluster level	County border	County border	County border	County border	County border

Notes: The table shows the baseline RD estimates are robust to alternative bandwidth (panel A), quadratic RD polynomial (panel B), and triangular kernel (panel C). The unit of observation is town. The sample includes the post-reform years 2010 and 2015. The regression discontinuity estimation is conducted at the town level along county borders with different treatment statuses on each side. The bandwidths in panel A are selected following Imbens and Kalyanaraman (2012) while the bandwidths in panel B and C are selected following Calonico, Cattaneo and Titiunik (2014). Treat is a dummy variable indicating towns located in counties with the survey teams deployed in 2005. Standard errors clustered at the county border level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A19: Effect on citizen attitudes - Alternative cohort groups

Dep. var.:	(1) Trust in local officials	(2) Trust in most people	(3) Eval. of govt. performance	(4) Eval. of others' health
Treat × 1990s cohort	0.308**	-0.008	0.060***	0.000
	(0.136)	(0.025)	(0.018)	(0.034)
County FE	X	X	X	X
Cohort FE	X	X	X	X
County controls × Cohort FE	X	X	X	X
Survey wave FE	X	X	X	X
Citizen controls	X	X	X	X
Cluster level	County	County	County	County
Observations	41,385	41,476	41,665	29,215
R-squared	0.067	0.066	0.057	0.084
Mean dep. var.	5.16	0.54	0.80	0.25

Notes: This table shows the effect of the reform on citizens' attitudes towards local governments, using an alternative definition of affected and unaffected cohorts. The unit of observation is citizen (three waves of survey data from the China Family Panel Studies, 2012, 2014 and 2016). Treat is a dummy variable indicating counties with the survey teams deployed in 2005. 1990s cohort is a dummy variable indicating citizens born in the 1990s. The estimation equation is a cohort difference-in-differences specification that utilizes two sources of variation: (1) treatment counties versus control counties; (2) affected cohorts versus unaffected cohorts within the same county. The latter source of variation is built on insights in the psychology and political science literature that citizens' political attitudes are most permeable during teenage years and keep stable since one's 30s (Wolfinger and Rosenstone, 1980; Krosnick and Alwin, 1989). As such, the 1990s cohort are defined as the affected group as they would be younger than 30 during the reform period, and thus, their political attitudes would be most permeable. Those born in or before the 1970s are the unaffected cohorts, as they would be older than 30 in the reform period and thus be unaffected by the reform. Standard errors clustered at the county level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A20: Effect on citizen attitudes - Controlling for media access

Dep. var.:	(1) Trust in local officials	(2) Trust in most people	(3) Eval. of govt. performance	(4) Eval. of others' health
Treat × 1980s cohort	-0.052	0.010	0.027	0.010
	(0.133)	(0.021)	(0.018)	(0.034)
Treat × 1990s cohort	0.447***	-0.001	0.054**	0.004
	(0.158)	(0.028)	(0.020)	(0.033)
County FE	X	X	X	X
Cohort FE	X	X	X	X
County controls × Cohort FE	X	X	X	X
Survey wave FE	X	X	X	X
Citizen controls	X	X	X	X
Citizen media access	X	X	X	X
Cluster level	County	County	County	County
Observations	17,163	17,166	17,194	11,999
R-squared	0.070	0.086	0.061	0.109
Mean dep. var.	4.84	0.57	0.82	0.25

Notes: This table shows the effect of the reform on citizens' attitudes towards local governments, controlling for citizens' media access. The unit of observation is citizen (three waves of survey data from the China Family Panel Studies, 2012, 2014 and 2016). Treat is a dummy variable indicating counties with the survey teams deployed in 2005. 1980s cohort and 1990s cohort are dummy variables indicating citizens born in the 1980s and the 1990s, respectively. The estimation equation is a cohort difference-in-differences specification (equation 14) that utilizes two sources of variation: (1) treatment counties versus control counties; (2) most affected cohorts versus less unaffected cohorts within the same county. The latter source of variation is built on insights in the psychology and political science literature that citizens' political attitudes are most permeable during teenage years and keep stable since one's 30s (Wolfinger and Rosenstone, 1980; Krosnick and Alwin, 1989). As such, the 1990s cohort are defined as the most affected group as they would be younger than 30 during the reform period, and thus, their political attitudes would be most permeable. The 1980s cohort are defined as the less affected group because some of them would be older than 30 during the reform period. The omitted group is those born in the 1970s, as they would be older than 30 in the reform period and thus be unaffected by the reform. To account for citizens' media access, I include three variables constructed from the survey regarding media access: (1) the number of days political news was accessed via television in the last week; (2) the number of days political news was accessed via Internet in the last week; (3) whether you have posted comments related to political issues and major national events on Internet in the past 12 months. Standard errors clustered at the county level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A21: Effect on citizen attitudes - IV estimates

Dep. var.:	(1) Trust in local officials	(2) Trust in most people	(3) Eval. of govt. performance	(4) Eval. of others' health
Treat × 1980s cohort	-0.027	-0.013	-0.001	-0.059
	(0.160)	(0.024)	(0.021)	(0.051)
Treat × 1990s cohort	0.532***	0.007	0.044*	-0.002
	(0.183)	(0.035)	(0.026)	(0.041)
County FE	X	X	X	X
Cohort FE	X	X	X	X
Province FE × Cohort FE	X	X	X	X
County controls x Cohort FE	X	X	X	X
Survey wave FE	X	X	X	X
Citizen controls	X	X	X	X
Cluster level	County	County	County	County
Observations	17,163	17,166	17,194	11,999
Effective *F*-statistic	44.46	44.24	44.43	45.49
Mean dep. var.	4.84	0.57	0.82	0.25

Notes: This table shows the IV estimates on the effect of the reform on citizens' attitudes towards local governments. The unit of observation is citizen (three waves of survey data from the China Family Panel Studies, 2012, 2014 and 2016). Treat is a dummy variable indicating counties with the survey teams deployed in 2005, and is instrumented by the randomly assigned rural survey teams in 1984. 1980s cohort and 1990s cohort are dummy variables indicating citizens born in the 1980s and the 1990s, respectively. The estimation equation is a cohort difference-in-differences specification (equation 14) that utilizes two sources of variation: (1) treatment counties versus control counties; (2) most affected cohorts versus less unaffected cohorts within the same county. The latter source of variation is built on insights in the psychology and political science literature that citizens' political attitudes are most permeable during teenage years and keep stable since one's 30s (Wolfinger and Rosenstone, 1980; Krosnick and Alwin, 1989). As such, the 1990s cohort are defined as the most affected group as they would be younger than 30 during the reform period, and thus, their political attitudes would be most permeable. The 1980s cohort are defined as the less affected group because some of them would be older than 30 during the reform period. The omitted group is those born in the 1970s, as they would be older than 30 in the reform period and thus be unaffected by the reform. The effective *F*-statistics are constructed following Montiel Olea and Pflueger (2013). Standard errors clustered at the county level are reported in parentheses. * p <0.1, ** p <0.05, *** p <0.01.

Table A22: Correcting for multiple hypothesis testing

Dep. var.:	(1) Standardized index of govt. policies	(2) IHS(Total loans)	(3) IHS(Firm entry)	(4) Trust in local officials	(5) Eval. of govt. performance
Treat $\times$ Post	0.474***	0.174	0.046*		
	(0.167)	(0.123)	(0.026)		
Raw p-value	[0.006]	[0.158]	[0.078]		
Sharpened q-value	{0.022}	{0.118}	{0.085}		
Treat $\times$ 1980s cohort				-0.051	0.026
				(0.129)	(0.018)
Raw p-value				[0.695]	[0.148]
Sharpened q-value				{0.226}	{0.118}
Treat $\times$ 1990s cohort				0.452***	0.052**
				(0.154)	(0.020)
				[0.004]	[0.012]
				{0.022}	{0.022}

Notes: This table adjusts for multiple hypothesis testing for the estimates on several aggregate outcomes used to measure local officials' development effort. To this end, I reproduce the results for several aggregate outcomes in this table (see the description of these results in Section 5), and report the sharpened q-values proposed by Anderson (2008) in braces to correct for multiple hypothesis testing. The raw p-values are reported in brackets for comparison.

Table A23: Fiscal transfers

	(1)	(2)	(3)
Dep. var.:		IHS(Fiscal Transfers)	
Transfer type:	All transfers	General transfers	Specialized transfer
Treat × Post2005	0.059	0.025	0.070
	(0.151)	(0.159)	(0.145)
County FE	X	X	X
Year FE	X	X	X
Cluster level	County	County	County
Observations	12,453	12,453	12,453
R-squared	0.529	0.513	0.538

Notes: This table examines whether the launch of the survey teams would increase fiscal transfers from upper-level governments. The unit of observation is county. The sample is from 2000 to 2007. Treat is a dummy variable indicating counties with the survey teams deployed in 2005. Post2005 is a dummy variable for years after 2005. The dependent variables are transformed by inverse hyperbolic sine (IHS) to reduce influences from the tails of the skewed outcome distributions (Bellemare and Wichman, 2020). Standard errors clustered at the county level are reported in parentheses. * p <0.1, ** p <0.05, *** p <0.01.

Table A24: Effect on policy diffusion

	(1)	(2)	(3)
Dep. var.:	Similarity of government work reports between county i and county j		
County pair type:	All types	Within province	Across province
Treat × Post	-0.000	-0.001	-0.002*
	(0.001)	(0.001)	(0.001)
County (i) × Year FE	X	X	X
County (j) × Year FE	X	X	X
County pair FE	X	X	X
Cluster level	two-way (i,j)	two-way (i,j)	two-way (i,j)
Observations	49,506	31,845	17,561
R-squared	0.977	0.979	0.987
Mean dep. var.	0.418	0.403	0.447

Notes: This table shows the effect of the reform on policy diffusion across counties. The unit of observation is county pair ij (with ij equivalent to ji). Treat is a dummy variable indicating county pairs in which both counties had a survey team deployed in 2005. Post is a dummy variable indicating years after the reform in 2009. The dependent variable is the pairwise textual similarity of government work reports, which is calculated following Kelly et al. (2021) and described in Appendix C. Column (1) considers all county pairs. Column (2) considers county pairs within the same province. Column (3) considers county pairs spanning different provinces. Standard errors that are two-way clustered by both counties in a pair are reported in parentheses. * p <0.1, ** p <0.05, *** p <0.01.

Table A25: Effect on government policies - social welfare and environmental protection

	(1)	(2)	(3)	(4)	(5)	(6)
Method:	Key words frequency		Random Forest score		SVM score	
Dep. var.:	Social welfare	Enviro. protection	Social welfare	Enviro. protection	Social welfare	Enviro. protection
Panel A: OLS estimates						
Treat × Post	0.000	-0.000	0.002	-0.001	0.000	-0.002
	(0.000)	(0.000)	(0.002)	(0.001)	(0.002)	(0.002)
R-squared	0.378	0.381	0.304	0.369	0.334	0.397
Mean dep. var.	0.002	0.001	0.015	0.018	0.018	0.022
Panel B: IV estimates						
Treat × Post	0.000	-0.000*	0.000	-0.003	-0.000	-0.003
	(0.000)	(0.000)	(0.002)	(0.002)	(0.002)	(0.003)
Effective *F*-statistic	74.06	74.06	74.06	74.06	74.06	74.06
Mean dep. var.	0.002	0.001	0.015	0.018	0.018	0.022
County FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
County controls × Post	X	X	X	X	X	X
Cluster level	County	County	County	County	County	County
Observations	883	883	883	883	883	883

Notes: This table shows the effect of the reform on government policies on social welfare and environmental protection. The unit of observation is county. The sample period is 2005-2018. The sample includes 97 counties. Treat is a dummy variable indicating counties with the survey teams deployed in 2005, and is instrumented by the randomly assigned rural survey teams in 1984 in panel B. Post is a dummy variable indicating years after the reform in 2009. Columns (1)-(2) measure policies using a simple key words frequency method. Columns (3)-(6) measure policies using supervised machine learning methods (Random Forest and Support Vector Machine). The detailed procedures for constructing these measures are described in Appendix C. The effective *F*-statistics in panel B are constructed following Montiel Olea and Pflueger (2013). Standard errors clustered at the county level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A26: Effect on local government debt

	(1)	(2)	(3)
Dep. var.:		IHS(Bond issuance)	
Treat	0.101	-0.019	-0.173
	(0.278)	(0.258)	(0.254)
County controls		X	X
Province FE			X
Observations	1,752	1,498	1,498
R-squared	0.002	0.244	0.183
Effective *F*-statistic	1,942	2,041	2,214

Notes: This table shows the IV estimates on effect of the reform on local government debt. The unit of observation is county. Treat is a dummy variable indicating counties with a survey team deployed in 2005, and is instrumented using the randomly assigned rural survey teams in 1984 as the instrument. The dependent variable is the total amount of bond issuance by local government financing vehicles (L-GFVs), which serves as a proxy for local government debt. The earliest issuance at the county level was in 2009. The dependent variables are highly skewed and are thus transformed by inverse hyperbolic sine (IHS) to reduce influences from the tails (Bellemare and Wichman, 2020). The effective *F*-statistics are constructed following Montiel Olea and Pflueger (2013). Robust standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A27: Effect on corruption convictions

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	# Corruption convictions					
Time period:	2012-2016			2015-2016		
Type of corruption:	All types	Bribery & Appropriation	Other types	All types	Bribery & Appropriation	Other types
Treat	0.343	0.242	0.101	0.188	0.072	0.117
	(0.301)	(0.231)	(0.142)	(0.221)	(0.165)	(0.120)
# Anti-corruption inspections	-0.067	-0.060	-0.007	-0.029	-0.037	0.008
	(0.060)	(0.045)	(0.029)	(0.044)	(0.032)	(0.025)
County controls	X	X	X	X	X	X
Province FE	X	X	X	X	X	X
Cluster level	County	County	County	County	County	County
Observations	1,498	1,498	1,498	1,498	1,498	1,498
R-squared	0.099	0.109	0.027	0.068	0.073	0.020
Effective *F*-statistic	2,204	2,204	2,204	2,204	2,204	2,204
Mean dep. var.	4.648	3.303	1.344	2.933	1.997	0.935

Notes: This table shows the IV estimates on the effect of the reform on corruption convictions. The unit of observation is county. Treat is a dummy variable indicating counties with the survey teams deployed in 2005, and is instrumented using the randomly assigned rural survey teams in 1984 as the instrument.. The dependent variable denotes the number of corruption convictions by type and period. This data contains 10,797 corruption convictions from 2005 to 2016, with a vast majority (10,788) happening after 2012 when China's anti-corruption campaigns began. The few convictions (9) before 2012 were likely caused by the lack of enforcement instead of less corruption, and are dropped from my analysis. The number of anti-corruption inspections are included to address concerns about differential enforcement. The effective F-statistics are constructed following Montiel Olea and Pflueger (2013). Robust standard errors are reported in parentheses. * p <0.1, ** p <0.05, *** p <0.01.

Table A28: Effect on corruption perception

	(1)	(2)
	OLS	IV
Dev. var.:	Corruption perception	
Treat × 1980s cohort	-0.201	-0.255
	(0.152)	(0.191)
Treat × 1990s cohort	-0.116	0.035
	(0.202)	(0.241)
County FE	X	X
Cohort FE	X	X
Province FE × Cohort FE		X
County controls × Cohort FE	X	X
Survey wave FE	X	X
Citizen controls	X	X
Cluster level	County	County
Observations	10,747	10,747
R-squared	0.103	0.038
Effective F-statistic		42.18
Mean dep. var.	4.85	4.85

Notes: This table shows the effect of the reform on citizens' perceived corruption about the government. The unit of observation is citizen (two waves of survey data from the China Family Panel Studies: 2014 and 2016). Treat is a dummy variable indicating counties with the survey teams deployed in 2005, and is instrumented by the randomly assigned rural survey teams in 1984 in column (2). 1980s cohort and 1990s cohort are dummy variables indicating citizens born in the 1980s and the 1990s, respectively. The estimation equation is a cohort difference-in-differences specification (equation 14) that utilizes two sources of variation: (1) treatment counties versus control counties; (2) most affected cohorts versus less unaffected cohorts within the same county. The latter source of variation is built on insights in the psychology and political science literature that citizens' political attitudes are most permeable during teenage years and keep stable since one's 30s (Wolfinger and Rosenstone, 1980; Krosnick and Alwin, 1989). As such, the 1990s cohort are defined as the most affected group as they would be younger than 30 during the reform period, and thus, their political attitudes would be most permeable. The 1980s cohort are defined as the less affected group because some of them would be older than 30 during the reform period. The omitted group is those born in the 1970s, as they would be older than 30 in the reform period and thus be unaffected by the reform. The effective F-statistics are constructed following Montiel Olea and Pflueger (2013). Standard errors used to construct the 90% confidence intervals, denoted by the spikes, are clustered at the county level. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.