Title
Three Essays on Local Governance in China

Author
Wang, Shaoda

Publication Date
2019

Peer reviewed|Thesis/dissertation

Three Essays on Local Governance in China

By

Shaoda Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Agricultural and Resource Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Elisabeth Sadoulet, Chair
Professor Noam Yuchtman
Professor William Reed Walker
Professor Jeremy Magruder

Spring 2019

Three Essays on Local Governance in China

Abstract

Three Essays on Local Governance in China

by

Shaoda Wang

Doctor of Philosophy in Agricultural and Resource Economics

University of California, Berkeley

Professor Elisabeth Sadoulet, Chair

This dissertation studies local governance in China. Guided by theoretical insights originated from the political economy and public economics literature, I exploit natural experiments and randomized field experiments to provide causal evidence on the role of bureaucratic incentives in the implementation of important public policies. The findings in this dissertation could deepen our understanding of the core incentives and constraints faced by Chinese local bureaucrats, and shed new lights on the design of more effective public policies.

In Chapter 1 (coauthored with Guojun He and Bing Zhang), we estimate the effect of environmental regulation on firm productivity using a spatial regression discontinuity design implicit in China's water quality monitoring system. Because water quality readings are important for political evaluations, and the monitoring stations only capture emissions from their upstream regions, local government officials are incentivized to enforce tighter environmental standards on firms immediately upstream of a monitoring station, rather than those immediately downstream. Exploiting this discontinuity in regulation stringency with novel firm-level geocoded emission and production datasets, we find that upstream polluting firms face a 27% reduction in Total Factor Productivity (TFP), and a 48% reduction in emission intensity, as compared to their downstream counterparts. We find that the discontinuity in TFP does not exist in non-polluting industries, only emerged after the government explicitly linked political promotion to water quality readings, and was entirely driven by prefecture cities with career-driven leaders. Linking the TFP estimate with the emission estimate, a back of the envelope calculation indicates that China's current water-pollution abatement target leads to an annual economic loss of more than 30 billion dollars.

Chapter 2 intends to provide causal evidence on fiscal competition and coordination between local governments. We first introduce a unique empirical setting created by the combination of two national programs in China, the "Rural School Consolidation (RSC)" program and the "Special Economic Zones (SEZs)" program, in which spatially adjacent townships have discontinuous incentives to either compete or coordinate with a neighboring SEZ. Exploiting such quasi-experimental variation with a spatial-discontinuity design, we find a series of empirical evidence on the existence of competition and

coordination between local governments, which can be formally rationalized by a simple model where asymmetric jurisdictions compete for mobile labor. Our results suggest that the strategic interactions among local governments play important roles in public goods provision and labor migration: during the RSC program (2001-2011), townships that coordinate with their neighboring SEZs closed an extra 15% of local primary schools, and lost an extra 7% of local population, as compared to townships that compete with their neighboring SEZs.

Chapter 3 is joint with Alain de Janvry, Guojun He, Elisabeth Sadoulet, and Qiong Zhang. This paper studies the personnel economics of Chinese grassroots bureaucrats. We notice that subjective performance evaluation is widely used by firms and governments to provide work incentives. However, delegating evaluation power to senior leadership could cause influence activities: agents might devote much efforts to please their supervisors, rather than focusing on productive tasks that benefit their organizations. We conduct a large-scale randomized field experiment among Chinese grassroots state employees and provide the first rigorous empirical evidence on the existence and implications of influence activities. We find that state employees are able to impose evaluator-specific influence to affect evaluation outcomes, and this process could be partly observed by their co-workers. Furthermore, introducing uncertainty in the identity of the evaluator, which discourages evaluator-specific influence activities, can significantly improve the work performance of state employees.

Table of contents

List of figures

List of tables

To my parents

To my wife

Acknowledgments

First and Foremost, I am deeply indebted to Professors Elisabeth Sadoulet and Alain de Janvry for their never-failing support since the very beginning of my academic pursuit. I first met Betty and Alain in 2012, when I was visiting Berkeley as an undergraduate student from Peking University. Since then, they taught me about development that made me decide to pursue a career in this field, offered me the chance to come back to ARE for graduate school, and spent countless hours supporting me in the past five years. One day when I have my own students, I want to be an advisor like Betty and Alain.

During my undergraduate exchange at Berkeley, I was also lucky to meet Professor Jeremy Magruder. Jeremy later encouraged me to come to ARE for graduate school, served as my orals exam chair and dissertation committee member, and generously provided detailed and insightful feedback on every project of mine. I really enjoyed all the helpful and delightful conversations with Jeremy over the years.

Since my first year in graduate school, I have been benefiting enormously from Professor Noam Yuchtman's intellectual support. In addition to being exceptionally sharp on details, Noam always urged me to think about the "big picture" and "good economics" behind each empirical exercise, which really influenced how I think about economics research in important ways. When I decided to do a "targeted search" for the job market, Noam went above and beyond to help me get great offers, and always had my best interest in mind. I feel truly lucky and proud to be Noam's student.

In my third year, when I started to get interested in environmental issues, I audited Professor Reed Walker's class and went to his office hours for advice. The encouragement and generous support I received from Reed fundamentally changed my career path: I started working at the intersection of development and environmental economics, which led to the opportunity of joining my dream school for my first job. I am deeply grateful for all the important things I learned from Reed, and I look forward to learning a lot more from him in the future.

In addition to my advisors mentioned above, many other professors at Berkeley have also been extremely generous to me with their time. I learned a lot from Professor Ernesto Dal Bo, both in his fascinating class and during many conversations throughout the years. I also very much enjoyed Professor Frederico Finan's political economy and development economics classes, which shaped my research interests when I was a second year student. I am really thankful to Professors Aprajit Mahajan, Ethan Ligon and Marco Gonzalez Navarro, who have provided me with great suggestions during the development workshop presentations and our private conversations. I also enjoyed taking Professor Michael Anderson's applied econometrics class, and thank him for allowing me to be the Graduate Student Instructor for that class later on. I also benefited from conversations with

Professors Thibault Fally, Meredith Fowlie, Larry Karp, Gordon Rausser, Joeseph Shapiro, Brian Wright, Gerard Roland, Edward Miguel, Peter Berck, and Leo Simon.

Of course I am also deeply grateful to my coauthors, Guojun He, Bing Zhang, Zenan Wang, Boxiao Zhang, Xiaoyang Ye, Ao Wang, and Qiong Zhang. I look forward to many more fruitful collaborations with them in the future. I also thank my peers at Berkeley for their help and friendship, an incomplete list includes Ellen Lin, Junyi Hou, Dave McLaughin, John Loeser, Kwabena Donkor, Itai Trilnick, Wei Lin, Manaswini Rao, Yang Xie, Weijia Li, Zhimin Li, Yueran Zhang, Xiangjun Feng, and Wenfeng Qiu. I am especially grateful to Sakib Mahmood for being a great friend, and for all the soccer games that we played together. Many faculties and peers from other institutions also offered me their generous help at different points, for which I am sincerely grateful, but the list of names is too long for me to get into great details here.

Finally, none of this would have been possible without the unconditional support from my family, especially my parents and my wife. I dedicate this dissertation to them.

1 Environmental Regulation and Firm Productivity in China: Estimates from a Regression Discontinuity Design

1.1 Introduction

The question of whether environmental regulation hinders factor productivity has long been important and controversial. On the one hand, neoclassical models suggest that environmental regulations will increase production costs, lead to a reallocation of labor and capital, and reduce the competitiveness of firms. On the other hand, proponents of environmental protection argue that stringent regulations provide incentives for polluters to invest in cleaner and less costly technologies to reduce pollution, which can in turn enhance productivity.[1] This debate entails particularly significant policy ramifications for rapidly-developing countries such as China and India, where the economies still rely heavily on manufacturing output in dirty industries, causing billions of people to suffer from extreme levels of water and air pollutions (Greenstone and Hanna, 2014; Ebenstein et al., 2017).

However, while there exists a large empirical literature investigating the relationship between environmental regulation and productivity (e,g, Jaffe et al. 1995; Berman and Bui, 2001; Greenstone, 2002; Greenstone et al., 2012), these works mostly focus on developed countries, and much less is known about the developing world, despite tremendous potential policy relevance. Developing countries may face a very different regulation-productivity tradeoff partly because they fall behind in industrial structures and technologies. But more importantly, developing countries tend to have weaker formal institutions, where local bureaucrats have drastically different incentives and constraints compared to those in the developed world.[2] Understanding the political economy of regulation is thus critical for the design of effective environmental programs (Acemoglu and Robinson, 2013; Greenstone and Jack, 2015).

In this study, we focus on China, the largest manufacturer and emitter in the world, where high-powered political incentives were leveraged to help alleviate pollution. Specifically, we exploit the political incentives implicit in China's surface water quality monitoring system, and estimate how tighter water emission controls affect the production and emission activities of Chinese manufacturing firms. Because water quality monitors can only pick up pollution information from upstream regions, and because the readings from these monitors are important for political evaluations, local government officials have particularly strong incentives to abate the emissions from upstream firms. As a result, within a small neighborhood around a water quality monitoring station, upstream firms typically face tighter environmental standards than downstream firms. By focusing on a narrow geographic band that only stretches a few kilometers upstream and downstream of each surface water monitoring station, we are able to causally identify the impacts of water quality controls on manufacturing productivity.

[1] Notably, Porter (1991) argues that, if one country adopts more stringent environmental standards than a competitor, firms in this country will invest more in clean innovations, which in turn will enhance the country's growth. Evidence in favor of Porter's hypothesis is summarized in a recent review by Ambec et al. (2013).

[2] It has been shown that local bureaucrats in the developing world have considerable discretions in public service delivery (Martinez-Bravo 2014; Munshi and Rosenzweig 2015; He and Wang, 2017).

Exploiting this spatial discontinuity in regulation stringency embedded in China's water quality monitoring system, we find that polluting firms in the near upstream of monitoring stations have a 27% lower TFP, and a 48% lower emission intensity, as compared to their near downstream counterparts. To put the magnitude in context, the TFP gap is on average equivalent to upstream polluting firms losing 2 years of productivity growth.[3] Importantly, we show that the discontinuity is not driven by the endogenous location choices of the monitoring stations, nor by the endogenous sorting of polluting firms around monitoring stations. Instead, a rich set of evidence suggests that the baseline findings are indeed driven by upstream firms receiving tighter regulation: (1) the upstream-downstream TFP gap only exists in polluting industries, and does not apply to non-polluting industries; (2) the discontinuity only emerged after the central government started to link water quality readings to political promotions; and (3) only firms within a few kilometers upstream are regulated, as emissions from the further upstream would dissipate quickly and have small influence on water quality readings.

In addition to the baseline economic impacts, we conduct a series of additional analyses to understand the political economy of environmental regulation in China, which has largely been a black box to both academia and the public. First, we find that upstream firms pay more emission fees and taxes than do downstream firms, even though they actually have lower levels of outputs and emissions. This implies that local governments use double standards in environmental regulation. Second, the baseline effect is completely driven by prefecture cities with politically motivated leaders, and there is no discontinuity at all when the leaders do not have promotion incentives. Third, we find that the discontinuity is more salient around "automatic" monitoring stations, whose data are less susceptible to local political influence, suggesting potential manipulation of environmental data among those traditional "manual" stations.

Our findings contribute to the ongoing debate on the economic costs of environmental regulation in several important ways. First, as mentioned, most studies to date have focused on developed countries (e.g., Jaffe et al., 1995; Henderson 1996; Becker and Henderson, 2000; Berman and Bui, 2001; Greenstone, 2002; Walker, 2011; Greenstone, List, and Syverson, 2012; Ryan, 2012; Kahn and Mansur, 2013; Walker, 2013). In this paper, we investigate China, the largest developing country in the world, and highlight the significant economic costs of environmental regulation in a rapidly-growing manufacturing economy. Moreover, most of the existing literature on environmental regulation focuses on air pollution, and the few exceptions that look at water regulation mostly only evaluate the environmental consequences of water regulation (Greenstone and Hanna, 2014; Keiser and Shapiro, 2018), leaving a gap in knowledge regarding the economic costs of water regulation, which our paper fills in.

Second, our paper adds to the growing literature on the political economy of pollution (List and Sturm, 2006; Burgess et al., 2012; Kahn et al., 2015; Lipscomb and Mobarak, 2017; Jia, 2017). Specifically, our analyses on the political economy of water quality monitoring shed light on how environmental regulations are implemented at the local level, which could guide future policy design in important ways. We find huge variations in the effectiveness of environmental regulation and the spatial allocation of regulatory burdens, which can largely be explained by regional differences in the incentives and constraints of local government officials. We also find that the regulation effect is stronger when improved technology makes it harder to manipulate water

[3] With the same dataset and same method for TFP estimation, Brandt et al. (2012) finds that the average TFP growth among Chinese manufacturing firms in 2005 was 14%.

readings, suggesting that local politicians have incentives to misreport environmental data in the absence of a precise monitoring practice.

Third, the detailed firm-level data also allow us to explore the channels through which firms are affected, and helps us understand how different types of firms respond to regulations. For instance, we show that upstream firms have to make substantially more investments in machineries (abatement capital) to cope with tighter regulation. Heterogeneous analyses reveal that the TFP loss is almost exclusively experienced by private Chinese firms, so tightening environmental regulations in the future is likely to damage the competitiveness of private Chinese firms rather than state-owned or foreign firms. Tests on sorting suggest that to avoid the large impact of regulation on productivity, upstream firms tend to relocate in the long run. These results imply a redistribution of production, income, environmental quality and social welfare between upstream and downstream regions. This study therefore also speaks to several lines of literature on the impacts of environmental regulation on production (Becker and Henderson, 2000), employment (Greenstone, 2002; Walker, 2011), plant location choice (List et al., 2003), income and total welfare (Ryan, 2012), and foreign direct investment (FDI) (Fredriksson et al., 2003; Hanna, 2010; Cai et al., 2016).

Finally, understanding firms' abatement costs and responses to regulation is critical for optimal policy design. Combining the TFP estimates with the emission estimates, we can calculate the economic costs of tightening water pollution regulations. We estimate that a 10% reduction in chemical oxygen demand (COD) emissions leads to a 2.49% decrease in TFP, and China's target of reducing total COD emissions by 10% between 2016 and 2020 would cause a total loss in industrial output value of 990 billion Chinese yuan (159 billion US dollars for five years) under current policy design and enforcement practices.

The rest of this paper is structured as follows. Section II describes the institutional background and research design. Section III discusses the data and presents descriptive statistics. Section IV presents the baseline results. Section V examines the channels, explores the political economy of environmental regulation, and tests whether emission measures also differ across the monitoring stations. Section VI interprets the results and benchmarks their economic significance. Section VII concludes the paper.

1.2 Research Design and Empirical Setup

A. Water Quality Monitoring and Water Pollution Controls in China

As the world's largest developing country, China faces a variety of pressing environmental challenges, including prevalent water and air pollution. According to the World Bank (2007), roughly 70 percent of China's rivers were polluted and contained water deemed unsafe for human consumption (at the time of that report). Poor surface water quality has driven policymakers to propose regulations to protect water bodies and reverse the process of degradation.

To gather surface water quality information, a national water quality monitoring system was established in the 1990s to monitor surface water quality in major river segments, lakes and reservoirs, known as the "National Environmental Quality Monitoring Network-Surface Water Monitoring System" (NEQMN-SWMS). At the beginning, the monitoring system was intended mainly for scientific rather than regulatory purposes, and most of the station-level monitoring data were kept confidential by the government. No strict emission abatement targets were set by the

Chinese government between 1990s to early 2000s because economic growth was considered the country's priority. As a result, along with China's rapid economic growth, the country witnessed severe degradation of its ecological systems.

In 2002, Hu Jintao became the new political leader of China, taking power from Jiang Zemin, and held office until 2012. Given the country's mounting environmental challenges, the new president started to emphasize the importance of seeking a balance between economic growth and environmental sustainability. Notably, in 2003, President Hu proposed the "Scientific Outlook of Development" (SOD),[4] which sought integrated sets of solutions to economic, environmental and social problems, opening an era of environmental regulation.

Responding to the SOD slogan, the Ministry of Environmental Protection (MEP) increased its efforts to resolve the water pollution issue. In 2003, the MEP issued an updated version of NEQMN-SWMS and spread the "Technical Specification Requirements for Monitoring of Surface Water and Wastewater" to local governments, explicitly highlighting the importance of surface water quality monitoring. The new policy documents imposed emission reduction targets for all the 419 state-controlled stations at the time, and also started a trend of automation of these stations. To introduce public monitoring of environmental quality, the water quality readings from all state-controlled stations also became available for the public and were published in various environmental yearbooks starting from 2003.

During President Hu's political regime, the importance of clean surface water was emphasized and the central government adopted a target-based abatement system to control environmental pollutants. In particular, during the 11th Five-Year Plan (2006–2010), the emission abatement targets included (but are not limited to): (1) reducing COD emissions by 10% (from 141.4 million tons in 2005 to 127.3 million tons in 2010), (2) reducing the percentage of monitored water sections failing to meet Grade V National Surface Water Quality Standards from 26.1% in 2005 to 22% by 2010, and (3) increasing the ratio of monitored water sections (of the seven main bodies of water in China) meeting Grade III National Surface Water Quality Standards from under 41% in 2005 to 43% by 2010.[5] With these targets, the central government then allocated binding abatement requirements to each province, and provincial governors were required to sign individual responsibility contracts with the central government, documenting their emission abatement plans in detail. Provincial governors further assigned strict abatement mandates to prefectures and counties and used local environmental performance along with other criteria to assess and promote local government officials. Water pollution control thus became an important political task for local government officials.

Because rivers flow from higher to lower elevation, water quality monitoring stations can only capture emissions from their upstream areas, but not from downstream areas. Under the new political regime, local officials would have strong incentives to enforce tighter environmental regulations in upstream regions than in downstream regions. We exploit this spatial discontinuity and estimate the causal impact of tighter water pollution regulation on productivity. Because the Chinese government did not enforce stringent industrial pollution controls until 2003, we expect that, if water quality monitoring indeed influences firm productivity, this effect should be weaker or non-existent before 2003, and become stronger afterward.

[4] SOD can also be translated as the "Scientific Development Concept" or the "Scientific Development Perspective."

[5] Source: http://www.mep.gov.cn/gzfw_13107/zcfg/fg/gwyfbdgfxwj/201605/t20160522_343144.shtml

B. Location Choice of Water Quality Monitoring Stations

The purpose of establishing a water quality monitoring network is to achieve a comprehensive understanding of the country's surface water quality. The monitoring system covers the country's major rivers, lakes, and reservoirs. A monitoring station should be set in a way that can is spatially representative to its neighborhood water bodies and can properly reflect changes in water pollutants over time. Consequently, the locations of the monitoring stations were chosen based mainly on hydrological considerations.

According to the MEP, the monitoring stations must be placed in rivers with steady flows, wide water surfaces, and stable river beds, and must avoid stagnant water areas, backwater areas, sewage outfalls, rapids and shallow water. The MEP also requires that monitoring stations be established to serve "long-term" purposes, ensuring that short-term needs (such as avoiding or targeting pollution from a specific region or a specific firm) cannot be accommodated.

An important feature of station placement is that the MEP requires monitoring stations to be built close to hydrological stations, which enables the government to combine hydrological parameters with water quality information. Most hydrological stations were built in the 1950s-1970s and are used to collect meteorological and hydrological data.

In this paper, we focus on the state-controlled surface water quality monitoring stations. State-controlled stations are established and supervised by the MEP and the State Council of China. The water quality readings from the state-controlled stations are reported directly to the MEP to ensure data quality. Yearly average water quality readings from the stations are reported in the environmental yearbooks and the central government used these data to assess the environmental performance of local governments.

Aside from state-controlled stations, there are also local water quality monitoring stations and special stations designed to monitor the emissions of major polluters. The special monitoring stations are placed immediately downstream from the polluter to monitor its environmental performance. We do not have data for these types of stations.

C. Research Design and Econometric Model

We exploit the spatial discontinuity in regulation stringency around water monitoring stations to estimate the causal effect of regulation on TFP. The distance between a firm and a monitoring station serves as the running variable. We examine whether firms located immediately upstream from the monitoring station have lower productivity than adjacent downstream firms. This empirical strategy is similar in spirit to recent works that also exploits the flow of pollution along rivers for identification (Kaiser and Shapiro, 2017; Lipscomb and Mobarak, 2017), but is novel in that it utilizes a unique spatial discontinuity setting around the monitoring stations and explores the political incentives behind the regulation.

Specifically, the identifying assumption of our research design is that, due to spatial adjacency, firms located immediately upstream and downstream of monitoring stations should be balanced *ex ante* along various dimensions but will differ from each other only because upstream firms become more tightly regulated. The discontinuity can be estimated by both parametric and non-parametric approaches. Gelman and Imbens (2017) show that the parametric RD approach, which uses a polynomial function of the running variable as a control in the regression, tends to generate RD estimates that are sensitive to the order of the polynomial and have some other undesirable statistical properties. As a result, estimators based on local linear regression or other smooth functions are often preferred, because they can assign larger weights to observations that are closer to the threshold and therefore can produce more accurate estimates. We thus focus on a local linear approach, which can be estimated by the following equation:

$$(1)\ TFP_{ijk} = \alpha_1 Down_{ijk} + \alpha_2 Dist_{ijk} + \alpha_3 Down_{ijk} Dist_{ijk} + u_j + v_k + \varepsilon_{ijk}$$

$$s.t.\ \ -h \leq Dist_{ijk} \leq h$$

where TFP_{ijk} is the total factor productivity of firm i in industry j around monitoring station k. $Down_{ijk}$ is an indicator variable that equals 1 if firm i (in industry j) is downstream from monitoring station k, and 0 otherwise. $Dist_{ijk}$ measures the distance between firm i and monitoring station k, and h is the bandwidth length (i.e., the acceptable distance from the discontinuity for sample inclusion).

To account for industry and location-specific TFP determinants in the non-parametric estimations, we control for industry and monitoring station fixed effects u_j and v_k. The model essentially compares upstream and downstream firms in the same industry around the same monitoring station. The estimation of this non-parametric RD model with fixed effects is implemented using the two-step approach suggested by Lee and Lemieux (2010), where industry and station fixed effects are absorbed by running an OLS regression of TFP on a set of industry and station-specific dummies, and then apply the non-parametric estimations on the residual TFP obtained from OLS estimation.[6]

The choice of the optimal bandwidth h involves balancing the conflicting goals of focusing comparisons near the monitoring stations, where the identification assumption is most likely to be satisfied, and providing a large enough sample for reliable estimation. In this study, we rely on an MSE-optimal bandwidth h proposed by Calonico, Cattaneo, and Titiunik (2014) and Calonico, Cattaneo, Farrell (forthcoming), and experiment with various kernel weighting functions to ensure robustness.

The standard error is clustered at the monitoring station level to deal with the potential spatial correlation of the error term, as suggested by Cameron and Miller (2015). We also try two-way clustering at both the industry and the monitoring station levels and get quantitatively similar standard errors.

As a way to check the robustness, we also estimate a parametric RD model:

$$(2)\qquad TFP_{ijk} = \alpha_1 Down_{ijk} + f\left(Dist_{ijk}\right) + Down_{ijk} f\left(Dist_{ijk}\right) + u_j + v_k + \varepsilon_{ijk}$$

where TFP_{ijk} is the total factor productivity of firm i in industry j around monitoring station k. $f\left(Dist_{ijk}\right)$ is a polynomial in distance between firm i in industry j and monitoring station k. The polynomial function is interacted with the treatment dummy to allow flexible functional form on both sides of the cutoff, and u_j and v_k are industry and station fixed effects.

[6] Lee and Lemieux (2010) argue that, if there is no violation of the RD assumption that unobservables are similar on both sides of the cutoff, using a residualized outcome variable is desirable because it improves the precision of estimates without violating the identification assumption.

1.3 Data and Summary Statistics

A. Data

Our analyses are based on several datasets that together provide comprehensive information on the socio-economic conditions of townships, the production and performance of industrial firms, and emissions from heavy polluters centered around the monitoring stations.

Water Quality Monitoring Stations

We collect data from water quality monitoring stations from surface water quality reports in various environmental yearbooks from 2003-2010, which include the China Environmental Yearbooks, China Environmental Statistical Yearbooks, and China Environmental Quality Statistical Yearbooks. Data available in more than two different sources are cross-validated. The number of state-controlled monitoring stations varied slightly between years in these reports, ranging from 400 to 500 stations. We geocoded all the water quality monitoring stations.

Annual Survey of Industrial Firms Database

Our firm-level TFP is calculated using data from the Annual Survey of Industrial Firms (ASIF) from 2000 to 2007. The ASIF data include all the private industrial enterprises with annual sales exceeding 5 million Chinese yuan and all the state-owned industrial enterprises (SOEs). The data are collected and maintained by the National Bureau of Statistics (NBS) and contain a rich set of information obtained from the accounting books of these firms, such as inputs, outputs, sales, taxes, and profits.

The detailed production information allows us to construct TFP measures for each firm in each year. There are several approaches to estimating firm-level TFP and each requires different assumptions (Van Biesebroeck, 2007). In this paper, we use the consistent semi-parametric estimator suggested by Olley and Pakes (1996). The Olley-Pakes method addresses the simultaneity and selection biases in estimating TFP and has been the most widely-used method for the investigation of Chinese firms' productivity in the literature (see for example, Brandt et al., 2012; Yang, 2015). Using Olley-Pakes TFP estimator ensures that our estimates can be compared with the previous ones. The details of estimating TFP using the Olley-Pakes method are discussed in Appendix A. For robustness checks, we also construct alternative TFP measures based on other estimators, such as the Ackerberg et al. (2015) approach and the naïve labor/capital productivity measures.[7]

The ASIF data have been used in many studies investigating Chinese firms. A well-known issue is that the data contain outliers. We follow standard procedures documented in the literature to clean the data. We first drop observations with missing key financial indicators or with negative values for value added, employment, and fixed capital stock. We then drop observations that apparently violate accounting principles: liquid assets, fixed assets, or net fixed assets larger than total assets; or current depreciation larger than cumulative depreciation. Finally, we trim the data

[7] Note that the TFP measure we construct only accounts for private factors of production. Environmental goods are not considered as inputs, which is sometimes done in the environmental economics literature.

by dropping observations with values of key variables outside the range of the 0.5^{th} to 99.5^{th} percentile.[8]

The ASIF data have detailed address information for sampled firms in each year. We geocode the location of the 952,376 firms that appeared in the sample and then compute precise distance measures between each firm and its closest water quality monitoring station.

Because our research design is fundamentally cross-sectional, in the baseline analysis, we collapse the multi-year data into a cross-section and apply the RD estimators to it. The interpretation of the coefficients is therefore an average effect that persists for years. To fully utilize the dynamic structure, however, we also apply non-parametric RD estimators to different years and examine how the discontinuity changes over time.

Environmental Survey and Reporting Database

To investigate whether water quality monitoring indeed reduces water-related emissions, we collect firm-level emission data from China's Environmental Survey and Reporting (ESR) database, which is managed by the MEP.

The ESR database is the most comprehensive environmental dataset in China that provides firm-level (polluting-source level) emissions for various pollutants. The ESR database monitors polluting activities of all major polluting sources, including heavily polluting industrial firms, hospitals, residential pollution discharging units, hazardous waste treatment plants and urban sewage treatment plants. In this study, we keep only the ESR firms that are in the same polluting industries as the ASIF firms.

The sampling criteria in the ESR database is based on the cumulative distribution of emissions in each county. Polluting sources are ranked based on their emission levels of different "criteria pollutants," and those jointly contributing to the top 85% of total emissions in a county are included in the database. In this study, we use ESR data between 2000 and 2007, the same period as the ASIF database.

During our sample period, the "criteria pollutants" changed over time. In 2000, only chemical oxygen demand (COD) emissions and sulfur dioxide (SO_2) were "criteria pollutants." Polluting sources included in the database were therefore chosen based on their contributions to COD emissions or SO_2 emissions. In 2007, ammonia nitrogen (NH_3) and NOx also became "criteria pollutants."

Among all the pollutants, COD is most relevant to this study. COD is a widely-used water quality indicator that measures the amount of oxygen required to oxidize soluble and particulate organic matter in water.[9] It assesses the effect of discharged wastewater on the water environment. Higher COD levels mean a greater amount of oxidizable organic material in the sample, which

[8] More details about the construction and cleaning processes of the ASIF data can be found in Hsieh and Klenow (2009), Songet et al. (2011), Yu (2015), and Huang et al. (forthcoming).

[9] For example, COD abatement is used by the central government of China as a key performance indicator to assess local government efforts in environmental protection. In the 10th and 11th Five-Year Plans (2001-2005 and 2006-2010), COD was used as a primary criterion (along with ammonia-nitrogen) to set national abatement targets and conduct performance appraisals.

reduces dissolved oxygen levels. A reduction in dissolved oxygen can lead to anaerobic conditions, which are deleterious to higher aquatic life forms.

We focus on COD emissions because COD is the first water-related “criteria” pollutant used by the MEP, and the government explicitly set a 10% abatement target for COD emissions in the 11th Five-Year Plan. We also corroborate the findings on COD emissions by looking at the amount of wastewater discharge.

Like the ASIF, this dataset also includes detailed address information. We therefore geocode all the ESR firms and compute their distances to the nearest monitoring sites. The dataset allows us to construct total emission levels and emission intensity measures (emission levels divided by total output value) for large polluters in each county.

Unlike the ASIF, however, we are less confident about the quality of ESR data. While later we will show that the results for emissions are consistent with our story, we must acknowledge that the quality of the ESR data is relatively poor. For instance, we find that there exist a large number of missing values and zeros in the ESR data, and location, industry, firm, and year fixed effects combined can only explain a small portion of firms’ emissions. These patterns seem suspicious to us. In fact, this data quality issue can be the reason that the government relies on the readings of monitoring stations, instead of the ESR data collected from polluting firms, to measure the effectiveness of pollution control, although the latter would be more powerful and better-targeted in a world without data manipulation. Given such, in subsequent analyses, we primarily focus on outcomes from the ASIF dataset and use the emission results as supportive evidence.

Township-level Socio-economic Data

The National Bureau of Statistics (NBS) conducts the “Township Conditions Survey (TCS)” on an annual basis. It is a longitudinal survey that collects township-level socio-economic data for all the townships in China. We have access to the TCS data for 20 provinces in 2002 and use the township-level data to assess similarities between upstream and downstream townships.

Geo-data

We obtained township-level GIS boundary data in 2010 from the Michigan China Data Center. We use GIS data of China’s water basin system from the Ministry of Water Resources. We use GIS elevation data to identify upstream and downstream relationships. These GIS datasets are then matched to our geocoded township and firm datasets.

B. Data Matching

The data we have compiled are, to our knowledge, the most comprehensive and disaggregated collection ever assembled on water pollution and firm-level economic and environmental performance in China. The matching process involves several steps and is illustrated in Figure 1.1.

First, we only keep river monitoring stations, because lake and reservoir stations would not allow us to identify the upstream-downstream relationship. Then, we put a layer of the water basin system on the township GIS map, and only keep townships that have at least one river passing through it. Then, using each monitoring station as a center, we draw a circle with 10 km radius. We overlay all the geocoded firms (from the ASIF dataset and the ESR dataset) on the map, and only keep those falling in a 10-km circle, and lying in a township with a river passing through it. This procedure allows us to identify all the firms that are relevant for our research design.

After that, we calculate each firm's distance to its nearest monitoring station. In some cases (mostly in the eastern coastal areas), the distribution of monitoring stations can be very dense and multiple tributaries or branch rivers merge into the trunk streams. As a result, some 10-km circles overlap with each other, making it difficult to identify upstream and downstream relationships (i.e., an adjacent upstream firm for one monitoring station can be in the adjacent downstream of another monitoring station). We therefore exclude these water monitoring stations from our dataset. In some less-developed regions (mainly in the Western areas), the distribution of large industrial firms is so sparse that the 10-km circles around monitoring stations contain no firms from the ASIF or ESR datasets. We also drop these monitoring stations from our sample. After these exclusions, we are able to use 161 water quality monitoring stations. The distribution of our sampled monitoring stations is represented in Figure 1.2.

For each firm kept in the sample, we project its location onto the nearest river basin, and extract the elevation of that projected point. Then, we compare this elevation to the elevation of the adjacent monitoring station, so that we could decide for each firm whether it is in the upstream or downstream of its adjacent monitoring station. In the end, our sample includes 19,150 unique ASIF firms and 9,888 ESR firms from 544 townships, located around 161 water quality monitoring stations.

We attempted to match the firms across the ASIF and ESR samples. However, because these two datasets use different sampling criteria and are managed by different government agencies (using different coding systems), we were only able to match 10% of the ASIF firms with the ESR firms. The matched sample is too small for us to draw any credible statistical inferences. In addition, as mentioned, we have concerns about ESR data quality. Therefore, in subsequent sections, we analyze these two datasets separately and mainly emphasize the results from the ASIF data.

C. Balance Checks

The underlying assumption for our RD design is that, in the absence of environmental regulation, upstream and downstream firms should be *ex ante* identical. Since environmental regulation may affect many production decisions, it is in general difficult to test this assumption using firm-level data, which primarily contains time-varying variables. The only two (arguably) time-invariant covariates in the ASIF data, i.e. "firm establishment date" and "firm ownership type", are indeed smoothly distributed around monitoring stations, as shown in Table 1.1 Panel A.

As discussed in Section II Part A, surface water regulation was not strictly enforced until President Hu Jintao came into power in 2003, thus water monitoring stations should not affect upstream firms in the early years of our data.[10] We thus compare upstream firms with downstream firms using pre-2003 data and test the differences in key variables such as value added, profit, employment, capital, intermediate input and tax. As can be seen from Panel B of Table 1.1, the vast majority of these covariates are balanced between the two groups. These results are consistent with our identifying assumption.

In addition to the balance tests using firm-level data, we also conduct balance tests using township level data, which include a rich set of variables that are important for firm production, to provide additional evidence that upstream and downstream regions are comparable in aspects other than water regulation. The results are summarized in Appendix Table 4.2. We see that basic township characteristics are balanced, including township area, arable area, distance to county

[10] The dynamic analysis of the RD results will be discussed in Section V.

center, whether the township is an old-region town, whether it is an ethnic minority town, the number of residents, and the number of administrative villages.[11] In Panel B, we test whether basic infrastructure measures are similar between upstream and downstream townships before 2003. Again, the length of roads, number of villages with road access, number of villages with electricity access, and number of villages with tap water access are similar between upstream and downstream areas before water regulation became a binding constraint. Finally, production requires labor. We examine whether human capital differs significantly between upstream and downstream townships prior to 2003. In the township data, we have two relevant variables: the number of primary schools and the number of students enrolled in primary schools. Again, we find no evidence that upstream townships differ from downstream townships in this regard.

The results in Table 1.1 and Appendix Table 4.2 are encouraging, as they indicate that upstream and downstream firms are well-balanced for both time-invariant characteristics and pre-2003 covariates, and these firms are located in townships that are highly comparable. While it is, of course, impossible to completely rule out the presence of unobserved factors discontinuously change across the monitoring stations, these balance checks lend additional credibility to our research design.

1.4 Results

A. Effects of Water Quality Monitoring on TFP

We begin the analysis by visualizing our main findings. Figure 1.3 plots log TFP (adsorbing station FE and industry FE) against distance to the corresponding monitoring station. Each dot represents the average log TFP for firms within a bin of distance; their 95% confidence intervals are also presented. A fitted function is then overlaid on the graph to illustrate the discontinuity at the monitoring stations.

We divide the firms in ASIF into polluting industries and non-polluting industries based on the definition of polluting industries used by the MEP (in terms of COD emissions).[12] In Panel A, we show the RD plot for residual log TFP in the polluting industries. We see a sharp change in TFP at precisely the locations where the water quality controls take effect. The TFP of upstream firms is significantly lower than that of downstream firms in polluting industries. In addition, it can be seen from Panel A that the treatment effects only apply to firms in the immediate upstream (<5km), rather than firms in the further upstream. This corresponds to the fact that surface water pollutions tend to dissipate over space, so emissions from the further upstream have limited impact on water quality readings. In contrast, in Panel B, we do not observe any comparable discontinuity in TFP in non-polluting industries.

Table 1.2 quantifies the graphical findings in Figure 1.3. Panel A presents the RD estimates without any controls, for both polluting and non-polluting industries. We see that polluting firms located immediately downstream from monitoring stations have higher TFP, but there is no TFP difference for the non-polluting firms. The estimates are not statistically significant because of large standard errors.

---

[11] An old region refers to a Communist Party's revolutionary base region. An administrative village is organized by one village committee and may include several natural villages.
[12] Details of the polluting and the non-polluting industries are summarized in Appendix Table 4.1.

Our sample covers 161 water quality monitoring stations in 34 manufacturing industries. A simple RD regression, as reported in Panel A, would compare upstream and downstream firms from different clusters (monitoring stations) and industries, creating noise in the statistical inference. To address this issue, we control for station fixed effects in Panel B, and control for both station and industry fixed effects in Panel C. By doing so, we effectively compare the TFP differences station by station and industry by industry, and then average the differences across stations and industries. Comparing the RD estimates in Panels B and C to Panel A, we see that the magnitudes of the estimated impacts are largely unchanged throughout different specifications. This is important because it suggests that station- and industry-specific characteristics, while being important determinants of firms' TFP, are uncorrelated with the treatment status. As we control for the fixed effects, the RD coefficients get more precisely estimated, and thus become more statistically significant.

The estimates for the non-polluting industries are close to zero and none of them are statistically significant. This suggests that our results are indeed driven by environmental regulation, rather than other confounding differences between upstream and downstream areas. For both sets of results, the RD estimates are robust to different choices of kernel functions.

In our preferred specifications, which account for both station and industry fixed effects, the estimated gap in log TFP between upstream and downstream firms ranges from 0.31 to 0.35 for the polluting industries. These estimates imply that the water quality monitoring has reduced upstream firms' TFP levels by 26.7% ($e^{-0.31}$-1) to 29.5% ($e^{-0.35}$-1). While a 27% change in TFP is certainly substantial, the magnitude is better interpreted in its specific context: a period during which China experienced unparalleled overall industrial TFP growth. According to Brandt et al. (2012), who use the same data and the same method for TFP estimation, the average TFP growth among the ASIF firms was 14% in 2005. Having that benchmark in mind, our RD estimates indicate that more stringent environmental regulation in the immediate upstream of monitoring stations effectively slowed down firm productivity growth by two years. As will be discussed in the mechanisms section, this is likely due to the fact that upstream polluting firms need to invest capital in abatement facilities, instead of in production.

B. TFP Effects by Firm Ownership, Size, and Firm Age

In Table 1.3, we explore whether the effect of water quality monitoring on TFP varies by ownership, firm size and firm age. In light of the findings reported in Table 1.2, we focus on residual TFP with station and industry fixed effects absorbed.

In Panel A, we estimate the RD by firm ownership type and find that the baseline TFP loss is driven mainly by private Chinese firms. Water quality monitoring has no significant impact on the TFP of state-owned enterprises (SOEs) and foreign firms. This result may reflect the fact that environmental regulations are not binding for SOEs or foreign firms as a practical matter; they generally have greater bargaining power over local governments and thus face less stringent enforcement. Another possible explanation is that SOEs and foreign firms generally have superior *ex-ante* environmental performance compared to private Chinese firms and therefore are not affected by tighter regulations. However, given the relatively small number of observations for SOEs and foreign firms in our sample, these sub-sample null results should be interpreted with caution.

In Panel B, we investigate the impact heterogeneity by firm size. In China, the government adopts a policy strategy called "Grasping the Large and Letting Go of the Small" ("*Zhua Da Fang*

Xiao"). "Grasping the large" means that policymakers mainly target large enterprises, while "letting go of the small" means that the government exerts less control over smaller enterprises. The phenomenon has been widely documented in the context of economic reforms and policy implementation (see, for example, Hsieh and Song, 2015). In environmental regulation, many policies are also designed in such a way that larger firms need to meet larger abatement targets.[13] We investigate if this phenomenon is true in our setting. We define small firms as having less than 50 employees and the rest are categorized as large firms. The results in Panel B show that the TFP impacts are statistically significant only for larger firms. In other words, the "Grasping the Large and Letting Go of the Small" strategy seems to be applied to the context of water quality regulations, too.

In Panel C, we compare the TFP loss by firm age. We are interested in whether old firms and young firms respond differently to water quality monitoring. We define new firms as firms born in or after 2003, when China's environmental regulations became stringent. We then estimate the discontinuities separately for old and young firms using post-2003 data. We find that the TFP loss caused by water quality monitoring is statistically significant for both old and young firms. This finding is inconsistent with the "grandfathering" phenomenon, in which new environmental policies are often designed or implemented in such a way that older firms can be exempted from tighter regulations, because the cost of retrofitting existing facilities is higher than that of building new sources with cleaner technology. In the context of China's water quality monitoring, both young and old upstream firms have been under tighter regulation than downstream firms since 2003. In addition, the magnitude of treatment effects for young firms established after the regulation became stringent in 2003 is comparable to that of old firms, suggesting that the selective locational choices of young firms are not driving the baseline findings.

C. Results by Year

The stringency of water quality regulations has changed substantially during our sample period. Specifically, in 2003, President Hu Jintao proposed the "Scientific Outlook of Development" initiative to address the pressing environmental challenges in China. In the same year, the MEP upgraded the surface water quality monitoring system. In addition, starting in 2006, COD abatement became a key indicator in evaluating local environmental performance, which were explicitly linked to political promotions.

We thus hypothesize that the TFP effect of water monitoring should be intensified after 2003 and 2006, respectively. In Figure 1.4, we provide RD estimates separately for each year. We find that the TFP differences between upstream and downstream firms exactly match the policy changes we have discussed. Specifically, the estimate is close to zero from 2000 to 2002, and becomes larger in 2003, the year President Hu took office. The effect becomes statistically significant starting in 2006, the first year of the 11th Five Year Plan. The corresponding results are summarized in Table 1.4.

The finding that the monitoring effect was close to zero and statistically insignificant prior to 2003 is consistent with the balance tests presented in Table 1.1, and further justifies our identifying assumption: in the absence of tighter water quality regulations, upstream and downstream firms around the same water quality monitoring station have similar levels of productivity.

[13] See, for example, "The Top 10,000 Energy-Consuming Enterprise Program," which requires only large firms to abate carbon emissions: http://www.ndrc.gov.cn/zcfb/zcfbtz/201112/t20111229_453569.html

The dynamic pattern of the RD coefficients is also re-assuring in terms of ruling out alternative explanations: to the extent that one thinks the baseline results are driven by confounding factors, such factors have to be specific not only to upstream vs. downstream firms, polluting vs. non-polluting industries, but also specific to the timing of two independent political events happened in China in 2003 and 2006, respectively.

D. Endogenous Location of Monitoring Stations

Our qualitative discussions on the rules of setting up monitoring stations, the balance tests of firm-level and township level variables, the finding that the discontinuity is only evident for polluting industries, and the result that the discontinuity only emerges after 2003, all suggest that the identifying assumption in our RD design is likely to hold.

Nevertheless, one may still be concerned about the endogenous location of monitoring stations. For instance, a politically connected polluting firm has strong incentives to lobby the local government, so that the monitoring station would be established upstream of that firm. If these connected firms also receive other forms of benefits from the government that could affect their productivity, such as subsidies or loans, our RD estimates would be biased.

In this section, we use an instrumental variable (IV) approach to directly address this concern. We exploit the fact that, when monitoring stations were set up, local governments typically attempted to locate them closer to existing hydrological stations, so that data, equipment and technicians could be shared in order to achieve economies of scale in water monitoring.

A hydrological station collects hydrological data such as water levels, flow velocity, flow direction, waves, sediment concentration, water temperature, and ice conditions, as well as data on meteorological conditions such as precipitation, evaporation, air temperature, humidity, air pressure and wind. Because hydrological stations were set up between the 1950s and 1970s (a period when China barely had any industrial pollutions at all), and because their locations were chosen based purely on hydrological considerations, these locations should be orthogonal to the future socio-economic conditions of their neighborhoods. All the hydrological stations were built and supervised by the Ministry of Water Resources (MWR), instead of the Ministry of Environmental Protection (MEP), and play no roles in collecting any measures of water pollution.

Therefore, one would expect that, except for inducing the establishment of monitoring stations, the existence of a hydrological station alone should have minimal impact on the production and emission behaviors of adjacent firms. Utilizing this "exclusion restriction," we adopt "whether a firm is in the near upstream area of a hydrological station" as an instrumental variable (IV) for "whether a firm is in the near upstream area of a monitoring station" and estimate a 2SLS to quantify the impacts of water quality monitoring on TFP.

Empirically, we estimate the following first-stage regression:

$$UpMoni_{ijk} = \alpha \cdot UpHydro_{ijk} + \lambda_j + \sigma_\mathrm{k} + \epsilon_{ijk} \tag{3}$$

where $UpMoni_{ij}$ is a dummy variable indicating whether firm i in industry j is in the near upstream area (10 km) of monitoring station k; $UpHydro_{ijk}$ is a dummy variable indicating whether firm i in industry j is in the near upstream area (10 km) of a hydro-station k; λ_j and σ_k represent industry and monitoring site fixed effects; and ϵ_{ijk} is the error term. We then estimate the second stage regression:

(4) $$TFP_{ijk} = \alpha \cdot Up\widehat{Moni}_{ijk} + \lambda_j + \sigma_k + \epsilon_{ijk}$$

where TFP_{ijk} is the TFP of firm i in industry j near the neighborhood of monitoring station k; $Up\widehat{Moni}_{ijk}$ is the predicted value from the first stage regression; λ_j and σ_k are industry and monitoring site fixed effects; and ϵ_{ijk} is the error term.

The regression results are presented in Table 1.5. We estimate the effects separately for firms in the polluting industries and for firms in the non-polluting industries. First, we find that the locations of hydrological stations can strongly predict the locations of water quality monitoring stations (Columns 1 and 3): if a firm is near the upstream of a hydrological station, it is also more likely to near the upstream of a monitoring station. The 2SLS estimates show that being in the near upstream of a water monitoring station decreases the TFP of a polluting firm by 0.35 logarithmic units (Column 2), but it does not affect the productivity of non-polluting firms (Column 4).

Note that the regression results in Table 1.5 are not readily comparable to those in Table 1.2, as these two approaches use very different sources of variation in the data and estimate different treatment effects with different identifying assumptions. The RD design estimates the average treatment effect at the cutoff, whereas the IV estimates the local average treatment effect for firms near a hydrological station. Nevertheless, the closeness of the magnitudes of the estimates between the two approaches (0.31 to 0.35 versus 0.35), and the consistent findings in both sets of results, provide additional support to the causal relationship between water quality monitoring and firm TFP.

E. Sorting of Polluting Firms

Environmental policies can affect firms' production plans and their location choices. In particular, the pollution haven hypothesis (PHH) posits that polluting capital would flow from places with more stringent environmental regulations to places with less stringent regulations. This issue is important because it will affect the interpretation of the RD estimates: if certain types of firms are systematically more likely to sort away from the upstream, then our baseline findings could be partly driven by sample selection, instead of intensive margin treatment effect.

In this section, we conduct a series of analysis to formally investigate whether there indeed exists significant sorting, and whether it affects the interpretation of our baseline findings. First, we follow Cattaneo et al. (2017a, 2007b) and conduct a data-driven manipulation test. As shown in Panel A of Figure 1.5, using the baseline sample (collapsed cross-section for 2000-2007), we find that there is no discontinuity in the distribution of polluting firms around the location of monitoring stations. The formal statistical tests are summarized in Appendix Table 4.3. This test suggests that there does not significant sorting in our baseline sample, so sorting could not explain the baseline RD results.

The intuition for the lack of sorting is the following: the firms enter into our ASIF dataset are generally large ones, for whom it is very costly and difficult to relocate. Even if they want to relocate, they need to buy a new piece of land from another local government, and build a new factory, before they can move the labor and capital to the new site. For most large manufacturing firms, this whole process of relocation could take years to happen. Recall that in Figure 1.4, we showed that environmental regulation only became a binding constraint after 2003, and the regulation effect only became significant after 2005, while the sample ends in 2007. So even if polluting firms decided to relocate immediately after regulation started to hurt them, the time window in our sample period is simply too short for that to happen, which explains why we could not detect any significant sorting in the data-driven manipulation test.

Fortunately, while we no longer have TFP information after 2007, the ASIF dataset still allows us to identify the location of polluting firms up until 2013, which is a decade after water quality monitoring started to matter for polluting firms. To verify the intuition on the lack of sorting in the short run, we take the ASIF data in 2013, and re-run the same set of density tests.[14] As shown in Panel B of Figure 1.5, in 2013, the density of the polluting firms indeed becomes slightly discontinuous at the cut-off: fewer polluting firms are located in the immediate upstream of monitoring stations. This finding suggests that more polluting firms leaving (or less firms emerging in) the upstream regions to avoid tighter regulation in the long run. The corresponding density tests results are reported in Panel B of Appendix Table 4.3.

Taken together, results from these two data-driven manipulation tests suggest that sorting might exist, but only in the long run, not in the short run. Therefore, sorting could not explain our baseline RD results in the short run.

Since the data-driven manipulation tests are essentially comparing the density of polluting firms around the cut-off, a potential concern is that while the overall density remains unchanged, different types of polluting firms might differentially enter\exit the upstream and downstream areas, creating a confounding difference in the composition of firms. To address this concern, we keep only a balanced panel of polluting firms during 2000-2007, and re-run the baseline RD for this sub-sample. Since these firms were already in the sample before regulation became relevant in 2003, and remained in the sample until 2007, the RD coefficient estimated using this subsample should solely reflect the "intensive margin treatment effect," rather than any form of "sample selection effect." As shown in Appendix Table 4.4, while the point estimates are noisier due to reduced sample size, the balanced panel sample still show consistent treatment effects with similar magnitudes to the baseline. Again, this exercise confirms that our main findings cannot be explained by the sorting of polluting firms.

F. Spillover Effects

Spillover effects, i.e. water quality monitoring somehow also affecting downstream firms, would not be a concern in a market with perfect competition, where there are many firms and the output market is unaffected by local environmental regulation. However, in a market with imperfect competition or more complicated structures, spillovers can exist. In our empirical setting, both positive and negative spillovers can emerge, depending on how upstream firms and downstream firms interact.

If industries are highly concentrated and their major producers are geographically clustered near the water quality monitoring stations, then imposing tighter environmental regulations on upstream firms would cause positive spillovers to downstream firms. The reasons are twofold. First, because upstream and downstream firms are the main producers and competitors in the market, increased production costs for the upstream firms will raise the market price of their products. Competing downstream firms will thus benefit because of this change in market conditions, not just because of the environmental enforcement affecting their upstream counterparts. Second, tighter environmental regulations also may cause inputs, both labor and capital, to move toward the

[14] The "value-added" variable is only available in the ASIF dataset until 2007; therefore, data in the later years could not be used for TFP analysis.

downstream firms. If more productive factors flow to downstream firms, their TFP will be higher for this reason as well.

Negative spillovers will emerge when clustered firms are collaborating instead of competing. This is particularly true if clustered firms are vertically integrated along the supply chain. If (geographically) upstream firms produce inputs for downstream firms, or *vice versa*, environmental regulations that increase upstream firms' marginal costs of production will also make downstream firms less competitive.

In the presence of spillovers, when we extrapolate these estimates to the whole country, the existence of a positive spillover effect will exaggerate the economic costs of regulation, while a negative effect will attenuate the estimated costs.

To assess whether or to what extent our findings are confounded by the potential spillover effects, we conduct a formal test using placebo downstream firms. Specifically, we first replace the actual downstream firms by their best matches from the sample of firms that are not in the neighbourhood of any monitoring stations, based on pre-2003 data. We then re-estimate the regression discontinuities for the matched firms using post-2003 data. These matched firms serve as placebo firms which are not affected by the potential spillovers between actual upstream and downstream firms. The intuition is that, if the spillover effects are insubstantial (downstream firms are not affected by monitoring stations), the placebo firms should have TFP similar to the actual downstream firms. Using placebo downstream firms should lead to results that are quantitatively similar to the baseline estimates.

In practice, we take the pre-2003 collapsed cross-sectional data and use a nearest neighbor matching strategy that finds the best matched firm from the pool of firms that are located outside the 10 km radius of the water monitoring stations for each downstream firm. These placebo downstream firms resemble the actual downstream firms in terms of TFP, industry type, and industrial output value before 2003. We then replace the actual downstream firms by the placebo firms in the post-2003 sample and estimate the regression discontinuities.

The results are reported in Table 1.6. Upstream firms have significantly lower TFP than placebo downstream firms, suggesting that the baseline findings are not driven by a positive spillover effect on the downstream firms. We focus our discussion on the RD estimates after station and industry fixed effects are absorbed, in Panel B. Compared with placebo downstream firms, upstream firms' log TFP is 0.48 to 0.61 units higher. These estimates are slightly larger than those in Table 1.3, but the differences are statistically indistinguishable. That implies that, if there may exist some spillover effect, this effect should be slightly negative. Consequently, the estimates in Table 1.2 will only understate the economic costs of water pollution regulation.

G. Robustness to Different Specifications

We check the robustness of our findings in Table 1.7. In Panel A, we re-estimate our models using a method proposed by Calonico, Cattaneo, and Titiunik (2014) in which local linear regression estimates can be "bias-corrected" for biases resulting from choice of bandwidth. They also suggest an alternative method for calculating standard errors that is more conservative than the conventional procedure. Using these alternative methods, we generate results that are qualitatively similar to the results featured in our main analysis.

In Panel B, we use alternative bandwidth selectors. The bandwidth chosen in our main analysis is based on one common MSE (Mean Square Error)-optimal bandwidth selector for both sides across the cutoff. We supplement this analysis with five other bandwidth selectors: (1) MSE-two: two different MSE-optimal bandwidth selectors (below and above the cutoff) for the RD treatment effect estimator; (2) MSE-sum: one common MSE-optimal bandwidth selector for the sum of regression estimates (as opposed to the difference thereof); (3) CER (coverage error rate)-optimal: one common CER-optimal bandwidth selector for the RD treatment effect estimator; (4) CER-two: two different CER-optimal bandwidth selectors (below and above the cutoff) for the RD treatment effect estimator, and (5) CER-sum: one common CER-optimal bandwidth selector for the sum of regression estimates (as opposed to the difference thereof).[15] The results remain the same regardless of the bandwidth selector used.

In Panel C, we conduct a placebo test using "fake" monitoring stations. We move the original stations upstream or downstream by 5 km or 10 km and re-estimate the RD models. We find that the discontinuity in TFP is only evident at actual monitoring stations and not at the fake stations.

In the Appendix, we present more robustness checks. In Appendix Table 4.7, we report the RD estimates using the parametric approach, Equation (2). We find quantitatively similar results: water quality monitoring decreases polluting firms' TFP but has no impact on non-polluting firms. However, the estimates from the parametric approach are more sensitive to the choice of the polynomial function form and inclusion of different samples. In Appendix Table 4.8, we use an alternative TFP measure suggested by Ackerberg et al. (2015) as the outcome variable, and again the results remain unchanged.

1.5 Channels: What Happened to the Upstream Firms?

A. Regulation and Firm Production

How do firms respond to tighter environmental regulations? In this section, we examine the channels through which environmental regulation affects firms' TFP. To rationalize the baseline findings and guide the following discussions, we present a theoretical framework to illustrate how environmental regulation can negatively affect TFP in Appendix B. In this model, firms need to use extra labor and capital to clean up emissions and the government can enforce tighter environmental regulation by increasing the emission tax. Facing a higher emission tax, firms need to hire more labor and capital for emission abatement, but these extra inputs do not directly contribute to output production. As a result, tighter environmental regulation will lead to a reduction in firms' TFP.

In Table 1.8, we estimate the impacts of water quality monitoring on several key variables and test whether these findings are consistent with our theoretical predictions. Panel A of Table 1.8 summarizes the results for output-related measures: value added and profit. Although the effects of water monitoring are statistically insignificant, we see a tendency that downstream firms earn more profit despite not producing more product. In Panel B of Table 1.8, we focus on labor, capital, and intermediate input. Labor is measured by the number of employees. We find that upstream firms hire more employees (16%) but the effect is statically insignificant. Upstream firms also tend to use slightly more intermediate input for production. Capital stock is measured by two different methods for robustness, following Yang (2015) and Brandt at al. (2012) respectively. We see that upstream firms have significantly larger capital stock compared to their downstream counterparts and the effects are statistically significant using different estimators. These results are consistent

[15] Please refer to Calonico, Cattaneo, Farrell (forthcoming) for technical details.

with the ubiquitous anecdotal evidence that polluting firms have to buy expensive equipment or abatement facilities to cope with tighter environmental regulation standards.

The interpretation of "increased abatement capital expenses in the upstream" could be corroborated using the firm-level emission dataset, which has a variable documenting the amount of abatement equipment owned by the firm. As shown in Appendix Table 4.5, upstream polluting firms indeed own significantly more abatement equipment than their downstream counterparts, consistent with the capital stock results.

In Panel C of Table 1.8, we further present results for naïve (reduced-form) productivity measures, calculated by dividing firms' value-added by labor or capital stock. Similarly, both labor and capital productivity are higher in downstream firms, with the impact on capital productivity being statistically significant. These results show that our baseline findings are robust to the use of simpler and more transparent productivity measures and reflect a real loss in firm productivity, rather than being mechanical to specific procedures of TFP construction.

In Panel D of Table 1.8, we test the Porter Hypothesis. The outcome of interest is firms' investments in research and development. The results show that tighter environmental regulation increases firms' investment in research and development, which is consistent with the Porter Hypothesis, but the result is statistically insignificant due to large standard errors.

The results in Table 1.8 suggest that the impacts of environmental regulation on TFP are manifested through multiple channels. The overall pattern confirms the predictions of our model: facing tighter environmental regulations, firms need to install expensive facilities (and potentially also hire more labor) to abate emission, leading to lower productivity.

B. Regulation and Emissions

The model in Appendix B also predicts that tighter environmental regulations will decrease both emission levels and emission intensity (emission per unit of output). In other words, upstream polluting firms are expected not only to reduce total emissions, but also to adopt cleaner technologies. This is consistent with the previous finding that upstream firms have larger capital stock. In this section, we formally examine the impacts of water quality monitoring on firm's emission and emission intensity.

We apply the same set of RD estimators to firm-level emission data (which is equivalent to polluting-source level data) from the ESR database. We examine four water pollution outcomes: (1) total amount of COD emitted, (2) COD emission intensity (total COD/total output value), (3) total amount of wastewater discharged, and (4) wastewater discharge intensity (total wastewater/total output value).

Table 1.10 reports the local linear RD estimates for the four outcomes. In Panel A, we can see that both COD emissions and COD emission intensity are higher for downstream firms, and most results are statistically significant at the 5% or 10% level. COD emissions by polluters immediately upstream from monitoring stations are 0.75-0.99 logarithmic units lower than those from firms immediately downstream. This implies that water quality monitoring reduces COD emission levels in upstream firms by 52.8%–62.8% ($e^{-0.75}$-1 to $e^{-0.99}$-1). For COD emission intensity, water quality monitoring reduces the COD emission intensity in upstream firms by 38.7%–49.3% ($e^{-0.49}$-1 to $e^{-0.68}$-1). In Panel B, we examine wastewater discharge. Downstream firms tend to discharge more wastewater but the results are statistically insignificant due to large standard errors. The results for wastewater discharge intensity, however, are statistically significant at the 5% or 10% level.

Combining both sets of results, we conclude that upstream firms emit less COD and wastewater overall and also produce fewer COD emissions or less wastewater per value of output (by adopting cleaner technologies), confirming the theoretical predictions.

Recall that the ESR database samples the most polluting firms in each county. Given that we focus on a small region around each monitoring station, many of the upstream and downstream firms are located within the same county. This causes a potential selective attrition problem because upstream firms facing tighter regulations tend to emit less and are thus less likely to be sampled in the ESR database compared to downstream firms. If such selection exists, our results in Table 1.10 will be underestimated, because the upstream firms that reduced the most pollution are no longer included in the sample. Thus, when we evaluate the environmental benefits of water monitoring, the estimates in Table 1.10 should be regarded as lower bounds.

C. Political Economy of Water Quality Monitoring

Our empirical analyses have shown that upstream firms' productivity is negatively affected by water quality monitoring. Our explanation is that, because water quality readings are politically important, local officials have incentives to enforce tighter regulations on upstream firms than on downstream firms. In this section, we explicitly explore such political-economic incentives behind water quality monitoring.

First, we document how upstream and downstream firms are treated differently by the government. In the ESR dataset, we find that upstream firms have substantially lower COD emission and wastewater discharge. In the ASIF dataset, we also have information on the waste discharge fees paid by each firm in 2004. If the government imposes a "fair" rule of punishing upstream and downstream firms for emissions, we should expect downstream firms to pay more than upstream firms, due to their higher emission levels. However, as shown in Panel A of Table 1.10, we find that upstream firms actually pay significantly more waste discharge fees to the government, despite that they emit much less. In other words, local governments are able to charge firms at differentiated emission fee rates, even though these firms are located close to each other and are within the same administrative jurisdiction. This double-standard phenomenon is not unique in our study and has been documented in some other settings as well. For example, Liu (2017) investigates China's tax reform and finds that local governments of China are able to collect additional informal taxes from certain firms for fiscal revenue. Fan et al. (2017) study China's value-added tax (VAT) system and show that firms located farther away from local tax agencies experience the largest increase in tax burden after VAT enforcement costs are brought down by a new information technology. Local governments in China have substantial discretion in the management of taxation and various fees.

Second, we examine the political incentives of local officials. As documented in the Chinese meritocracy literature, China has an implicit rule that a prefecture-level governor cannot be promoted to a higher level if his/her age reaches 57 (Xi et al., 2017). This creates a discontinuous drop in political incentives at the age of 56. To test whether the TFP effects of water quality monitoring can be explained by political incentives, we digitize the résumés of every prefectural party secretary (the highest-ranked political leader in a prefectural city) between 2000 and 2007. We define a leader as "having strong political incentives" if he/she is younger than 56 in a given year, and "having weak political incentives" otherwise. We then assign a monitoring station either to an "incentivized" or "un-incentivized" party secretary in a given year and analyze two subsamples based on whether the monitoring station is under the governance of an "incentivized" leader in a particular year. The RD results are summarized in Panel B of Table 1.9.

We find that, when the prefectural city leader has strong political incentives, water quality monitoring has a statistically significant impact on upstream firms' TFP. The estimated impacts range from 0.57 to 0.66 using different kernel functions and are nearly twice as large as the baseline results in Table 1.2. In sharp contrast, when the prefecture city leader has weak promotion incentives, the TFP gap remains precisely at zero in all specifications. These results imply that the TFP discontinuity across the monitoring stations is driven by the political incentives of local officials.[16]

Third, despite the fact that state-controlled monitoring stations are established and run by the central government, it is still possible that local officials can exert their administrative powers to influence the water quality monitoring. Our concern is that, if local governments can manipulate water quality readings, they may be less incentivized to regulate upstream firms' emissions. These is evidence that air pollution data has been manipulated at the margin in some Chinese cities because air quality is important for political evaluation (Ghanem and Zhang, 2014).

To test this hypothesis, we estimate the RD separately for two types of monitoring stations: automatic stations and manual stations. Automatic stations conduct all water quality tests automatically and report the data directly to the central government, while manual stations require technicians to conduct the tests manually.[17] Because it is difficult for local governments to manipulate data from the automatic stations, we expect a larger TFP gap around automatic stations.

Panel C of Table 1.9 reports the findings: while we see an upstream-downstream TFP gap for both types of stations, this effect is much larger for automatic stations (almost three times larger). As the sample size shrinks substantially, most RD estimates are statistically significant only at the 10 percent level.

D. Enforcement in Practice

The results in the previous subsections are instrumental for the understanding of China's environmental regulations. In particular, we show that similar firms adjacent to each other can live in dramatically different regulatory environments, leading to different firm behaviors and diverging productivity.

To confirm these findings with qualitative materials, we conducted multiple field trips and interviewed dozens of firm managers and technicians working in the monitoring stations. Our discussions reveal that, not only do upstream firms bear larger burdens of emission fees, as shown in our empirical analyses, they also face a variety of command and control regulations that cannot be easily quantified in our data. For example, in one city, we learned that firms' production could be abruptly restricted or even suspended by the local government in order to improve water quality readings. In China's most polluted river basin, Huai River basin, environmental inspectors were

[16] In an alternative specification, we use the panel dataset, and exploit the age change from 56 to 57, holding the leader fixed. The main results still go through with this much more restrictive specification, as shown in Appendix Table 4.6.

[17] Most stations were manual in the 1990s and early 2000s, but these were gradually replaced by automatic stations, in order to improve the accuracy of water quality reporting. Weekly water quality reports from the automatic stations are posted by the MEP at http://datacenter.mep.gov.cn/index and real-time water quality readings can be accessed at http://online.watertest.com.cn/help.aspx.

placed in the polluting firms from time to time to ensure their compliance with environmental standards. These inspectors visited upstream firms more frequently because they knew these firms had large impacts on water quality readings. In some extreme cases, if certain firms do not comply with the regulation, electricity and natural gas supply could even be cut in order to meet the city's environmental abatement target.

Regulatory documents from local governments tell a similar story. Urgent orders were issued when local governments realized water quality readings might fail to meet higher-level government policy targets. A recent example of such an order, which attracted wide media attention in China, is presented in Appendix C. In this example, Kunshan city in Jiangsu Province required 270 manufacturing firms to suspend their production in order to improve water quality. Sluice gates along the rivers were closed and the pumping facilities were shut down so that no wastewater could be discharged into the rivers, even after abatement treatment. Special investigators were sent to the plants to enforce the production suspension policy. While some of these command and control policies cannot be quantitatively analyzed in our empirical analysis, they do help explain why China's water quality monitoring can be so effective in reducing water pollution and has such a significant impact on upstream firms' productivity.

1.6 Economic Significance

A. Economic Costs under Various Scenarios

Our baseline model estimates that water quality monitoring has caused an average loss in TFP of 0.31 logarithmic units for polluting firms (as shown in Panel B of Table 1.2), equivalent to a 26.7% drop. To translate this TFP loss into monetary value, one may ask what would happen if all of China enforced regulatory standards as stringent as those faced upstream. The total industrial output value (total revenue) from the polluting firms was about 11 trillion Chinese yuan (1,380 billion USD) in 2006.[18] If all these firms were subject to water quality monitoring regulations as stringent as those faced by the upstream firms in our empirical setting, the total annual loss in output value would exceed 4.0 trillion Chinese yuan (502 billion US dollars) based on 2006 industrial output value.[19]

However, the regulations faced by upstream firms may be too stringent to apply to all the other firms in the country. A more informative counterfactual would be to determine the TFP loss and economic costs associated with a given amount of emission abatement. Recall that all the firms in the ESR database together contribute 85% of China's total emissions, and all of them are local large emitters regardless of industry or revenues.

Because we are unable to match the ESR firms with ASIF firms, we cannot directly link the TFP estimates with COD estimates without imposing additional assumptions. The TFP and COD effects of water monitoring we estimated in previous tables essentially are the following:

(5) $$\mathrm{TFP}_{\mathrm{ATE}}|\mathrm{Revenue} \geq 5 \text{ million} = E(\mathrm{TFP}_1 - \mathrm{TFP}_0|\mathrm{Revenue} \geq 5 \text{ million})$$

[18] We use the 2006 exchange rate of 1:7.97.

[19] We compute the difference between the counterfactual output of 14,973.7 billion Chinese yuan (calculated by 10975.7/(1-26.7%)) and the observed output of 10,975.7 billion Chinese yuan in the polluting industries in 2006. The calculations for other parts follow the same method.

(6) $\mathrm{COD_{ATE}|COD \geq\ x = E(COD_1 - COD_0|COD \geq\ x)}$

where $\mathrm{TFP_{ATE}|Revenue \geq 5\ million}$ is the average treatment effect of water quality monitoring on TFP for firms with annual revenues over 5 million yuan, and $\mathrm{COD_{ATE}|COD \geq\ x}$ is the average treatment effect of monitoring on emitters that produce COD pollution more than a given threshold x. $\mathrm{TFP_1}$ is the TFP for downstream firms, and $\mathrm{TFP_0}$ is the TFP for upstream firms.

The average treatment effects on TFP and COD over the entire distribution are:

(7) $\mathrm{TFP_{ATE} = Prob(Revenue \geq 5\ million) \cdot TFP_{ATE}|Revenue \geq 5\ million\ + Prob(Revenue < 5\ million) \cdot TFP_{ATE}|Revenue < 5\ million}$

(8) $\mathrm{COD_{ATE} = Prob(COD \geq x) \cdot COD_{ATE}|COD \geq x\ + Prob(COD < x) \cdot COD_{ATE}|COD < x}$

where the probabilities could be written as the share of firms appearing in each sample:

$\mathrm{Prob(Revenue \geq 5\ million)} = \frac{N_{ASIF}}{N}$, $\mathrm{Prob(Revenue < 5\ million)} = 1 - \frac{N_{ASIF}}{N}$;

$\mathrm{Prob(COD \geq x)} = \frac{N_{ESR}}{N}$, $\mathrm{Prob(COD < x)} = 1 - \frac{N_{ESR}}{N}$.

While we cannot directly estimate "$\mathrm{TFP_{ATE}|Revenue < 5\ million}$" and "$\mathrm{COD_{ATE}|COD < x}$" in the data, we attempt to back them out by extrapolating the intra-sample heterogeneous treatment effects on TFP and COD.

In Table 1.11, we estimate the heterogeneous treatment effects of water quality monitoring on TFP with respect to firms' revenues, and the heterogeneous treatment effects of water quality monitoring on COD emission intensity with respect to firms' total COD emissions using the polynomial RD approach.[20] The revenue heterogeneity is estimated by using the polynomial RD approach with an interaction term between the downstream dummy and firms' revenue (log). We use the specification in Column 1 of Table 4.4 as our preferred parametric specification because it generates the closest RD estimates to the non-parametric RD estimates. To allow for non-linear heterogeneity, we also include quadratic and cubic interactions in the regressions. Based on the regression results, we then predict the estimated impacts at different levels of revenues and summarize the results in Panel A. We find that the TFP effect is substantially larger for larger firms and non-existent for smaller firms. The effects of water quality monitoring on TFP for the smallest 20% of firms (among all the firms with an annual revenue of above 5 million Chinese yuan) become negligible. The results are the same if we use quadratic or cubic heterogeneity. In Panel B, we conduct a similar analysis for COD emission intensity and check whether the effect of monitoring varies across different polluting sources. We find the same pattern: larger emitters are strongly affected by water quality monitoring, while the treatment effect becomes essentially zero for the smallest 20% of emitters in the ESR sample. These findings again confirm that the "Grasping the Large and Letting Go of the Small" strategy discussed in section IV B is applied to the context of water quality regulations.

[20] Ideally, we should also apply the non-parametric RD estimates to different sub-groups of firms and estimate the heterogeneity separately for each sub-group. However, doing so significantly reduces the sample size of each group and we would not have strong enough statistical power to make a reliable inference. In Appendix Table 4.6, we divide the sample into only two groups and find results that are largely consistent with Table 1.11: the impacts of water quality monitoring are primarily experienced by larger firms or emitters and are negligible for their smaller counterparts.

In addition, there is an "exit" variable in the ASIF database documenting whether a firm is excluded from the sample in the following year. A firm that earns less than 5 million Chinese yuan in a particular year, based on the sampling criteria, is dropped from ("exits") the database the next year. This outcome provides additional information on whether water quality monitoring affects firms at the margin. In Appendix Table 4.7, we find that the probability of exiting the ASIF database is not affected by water quality monitoring. This finding again shows that monitoring does not affect smaller firms at the margin.

Given these findings, if we assume that water quality monitoring does not increase the TFP or emission levels of upstream firms, and that the size of the treatment effect on TFP or emissions is a well-behaved function with respect to revenue or emissions, then we can make the following extrapolations:

$$\text{TFP}_{\text{ATE}}|\text{Revenue} < 5\text{ million} = 0 \quad (9)$$

$$\text{COD}_{\text{ATE}}|\text{COD} < \text{x} = 0$$

Intuitively, as the smallest producers and emitters in our ASIF or ESR dataset already have zero treatment effects, the even smaller producers and emitters (those excluded from the ASIF/ESR dataset) also should have zero treatment effects. We can therefore simplify Equations 7 and 8 to the following:

$$MRS = \frac{TFP_{ATE}}{COD_{ATE}} = \frac{N_{ASIF}}{N_{ESR}} \cdot \frac{\text{TFP}_{\text{ATE}}|\text{Revenue} \geq 5\text{ million}}{\text{COD}_{\text{ATE}}|\text{COD} \geq \text{x}} \quad (10)$$

The sample we use for estimation includes 6,581 firms in polluting industries from the ASIF database and 9,888 polluters from the ESR database. Using this equation, we can calculate the economic costs of water pollution abatement.

In Table 1.12, we compute the economic costs for various scenarios. For easy reference, Panel A reproduces the key results from Tables 1.2 and 1.10, and Panel B calculates the economic costs. We focus on the estimates in Column 1 because they produce modest values across all specifications. We first focus on COD emissions. Water quality monitoring reduces COD emissions by 0.83 logarithmic units and decreases TFP by 0.31 logarithmic units. A 10% change in total COD emissions causes a 2.49% change in TFP levels in the polluting industries.[21] Using alternative specifications produces slightly different results, which are reported in Columns 2–4. Similar interpretations can also be applied to COD emission intensity. In Column 1, an upstream firm's COD emission intensity is about 0.55 logarithmic units lower than that of a downstream firm. This means a 10% change in COD emission intensity causes a drop in TFP by about 3.75%. Other combinations create slight variations, as summarized in Columns 2–4.

During China's 11th Five-Year Plan, total COD emissions were reduced by 12.45% from 2006 to 2010 (with the target being 10%). If we attribute the entire COD reduction from 2006 to 2010 to the polluting industries, then this 12.45% abatement in COD emissions would cause a total output loss worth 352 billion Chinese yuan (44.2 billion USD) in the polluting industries, based

[21] The way we interpret this relationship is analogous to the Wald estimator in the two-stage setting, except that we do not have a readily available tool to combine the two stages from two different samples non-parametrically and we need to adjust for sample size. Water quality monitoring reduced COD emissions by 0.83 logarithmic units and TFP by 0.31 logarithmic units, so a 10% change in COD emissions will lead to a (6,581/9,888)*(0.31/0.83)*10% (= 2.49%) change in TFP.

on 2006 industrial output values.[22] The annual reduction in COD emissions between 2006 and 2010 was roughly 2.5%, equivalent to an annual loss of 69 billion Chinese yuan (8.7 billion US dollars) in gross industrial output value per year using 2006 Chinese yuan.

In 2015, the gross output value (of firms above designated size) in China exceeded 110 trillion Chinese yuan, and about 35% of output value (38.8 trillion Chinese yuan) is contributed by the polluting industries. The central government aims to reduce COD emissions by another 10% during the 13th Five-Year Plan, from 2016 to 2020. Applying our estimates to the 2015 data, we can infer that the total output loss would be around 990 billion Chinese yuan (159 billion US dollars) under current monitoring and enforcement practices.[23] Other specifications generate slightly different estimates, ranging from 936 to 1,099 billion Chinese yuan (150.5 to 176.7 billion USD).

Note that we make two strong assumptions in calculating the economic costs. First, we implicitly assume the marginal cost of abatement is linear so that the large and small emission reductions have proportional impacts on productivity. Second, we assume that the estimates in Table 1.10 are reliable and potential data manipulation would not significantly change our estimates.

B. Potential Sources of Bias

There are several reasons why the estimates in Table 1.12 may understate the true economic costs of China's water pollution controls. First, we use a conservative estimate of the effect of monitoring on TFP in our calculations. In fact, as shown in Table 1.3, the TFP loss due to water quality monitoring has increased from 0.31 to 0.40 since 2003. If we use these larger TFP estimates, the associated economic costs will increase.

Second, although we provide evidence that the smaller firms or emitters in our data are not affected by water quality monitoring, the assumption that water quality monitoring does not affect even smaller (unobserved) firms or emitters at all may still be violated. Shutting down very small polluters can be a feasible policy for some local governments to enforce tighter environmental standards. The TFP loss due to shutdown cannot be captured in our estimation.

Third, the distinction between polluting and non-polluting industries is based on two- to three-digit industrial codes. This distinction does not rule out the possibility that some firms in the non-polluting industries may also emit pollutants and are therefore regulated by local governments. If this is the case, the estimated TFP and economic loss are understated.

Fourth, some regions have a high density of monitoring stations as well as multiple tributaries along the main streams. These monitoring stations are excluded from our sample because we cannot credibly identify the upstream and downstream townships. If there are more monitoring stations in more polluted regions, some of the most polluted regions and firms are excluded from our sample; if environmental regulations are even more aggressively enforced in more polluted regions, the TFP loss in these regions could be even larger.

[22] We estimate that a 10% change in total COD emissions will cause a 2.49% change in TFP, which implies that a 12.5% change in total COD emissions will cause a 3.11% change in TFP. We then compute the difference between the counterfactual output of 11,328 (10,975.7/(1-3.11%)) billion and the observed output of 10975.7 billion in 2006. The calculations for other parts follow the same method.

[23] We use the 2015 exchange rate of 1 USD to 6.22 Chinese yuan.

Finally, we only compute the direct economic costs caused by TFP loss. Previous research has shown that tighter environmental regulation can also cause unemployment, firm relocation, and worker migration, and can change the flow of investment. These indirect costs are non-trivial and should be considered when calculating the overall economic costs of environmental regulations.

1.7 Conclusion

As the income levels of the Chinese people rise, the country starts to face a stark tradeoff between preserving high environmental quality and sustaining robust economic growth. This paper is the first study to credibly estimate the impacts of environmental regulations on Chinese manufacturing firms and provides a timely assessment of the economic costs of China's water pollution control policies. We exploit a regression discontinuity design based on the upstream-downstream relationship of water quality monitoring stations in China and find that tighter water quality regulations lead to significant TFP loss for firms located upstream from monitoring stations. This is the case for firms in polluting industries; such a discontinuity is not observed for firms in non-polluting industries.

We estimate that water quality monitoring reduces TFP levels by 26.7% in firms located immediately upstream from monitoring stations. This TFP loss is driven mainly by private Chinese firms instead of state-owned or foreign firms. A closer examination of the TFP effect by year reveals that the impacts of water quality controls have been greater in more recent years, consistent with the fact that environmental regulations in China have tightened over the past decade.

We also investigate the impacts of water quality monitoring on emissions. Using another firm-level dataset, we find that, at the extensive margin, upstream firms emit substantially (52.8%–62.8%) less COD and industrial wastewater than downstream firms; and, at the intensive margin, upstream firms adopt cleaner technology and emit less pollution per value of output (38.7%–49.3%).

Combining both sets of estimates, we calculate the economic costs of China's water pollution control policies. We estimate that a 10% abatement in COD emissions and COD emission intensity can lead to a 2.35%–2.75% and 3.43%–4.21% drop in a polluting firm's TFP respectively. These estimates imply that China's efforts in reducing COD emissions from 2016 to 2020 would cause a total loss in output of 936 to 1,099 billion Chinese yuan (150.5 to 176.7 billion US dollars) in the polluting industries, at least if current monitoring and enforcement practices remain unchanged.

Overall, our findings highlight the negative impacts of environmental regulations on productivity. The estimated efficiency loss is substantial; so high environmental quality comes at high economic cost. This is particularly salient for fast-growing economies that rely heavily on manufacturing.

Political incentives are fundamental to understanding China's environmental regulation. We show that the effect of environmental regulation depends on local officials' chances of promotion and local government's power to manipulate environmental data. Specifically, the TFP difference between upstream and downstream firms becomes twice as large when the city leader has a greater probability of promotion, and it approaches zero when the city leader cannot be promoted. The effect of water quality monitoring on TFP is substantially larger for stations that automatically test and report water quality to the central government.

Our findings also demonstrate that environmental regulations have profound distributional consequences. In the context of water quality monitoring, emission controls in upstream regions will improve the water environment in downstream regions. Upstream firms abate emissions and earn reduced profits, while downstream regions enjoy both higher environmental quality and more rapid economic growth. In the long run, these effects may imply a spatial redistribution of economic activity, population and social welfare.

Nevertheless, our findings do not answer the broader question of whether China's current environmental regulation standards are too aggressive or too lenient, because we do not know Chinese people's willingness to pay for cleaner surface water. After all, little research has been conducted on the socio-economic costs of water pollution in China.[24] To what extent environmental regulations should be designed and enforced, especially in developing countries that rely heavily on manufacturing industries, remains an underexplored research area.

We conclude by pointing out some limitations of this study and offering directions for future research. First, the estimates in this paper are derived in a partial equilibrium framework. We focus on a unique setting that affects only a small set of firms. Large-scale regulation will affect aggregate output and input markets, and our estimates should be interpreted with caution when used to evaluate large-scale environmental policies. Second, our sample covers a relatively short period of time, while firms might be able to better adjust investment and production in the long run. With the growing availability of firm-level longitudinal data, investigating how firms respond to regulation over long periods of time will be an important area for future research. Relatedly, sorting and its subsequent welfare implications are important for long-term impact assessment. Finally, with the expectation of increasingly tighter environmental regulations in China, entrepreneurs and investors may choose to develop businesses in non-polluting industries. Tighter environmental regulations in the polluting industries may create externalities affecting non-polluting industries, and there is a lack of rigorous empirical studies to quantify the impacts of such spillover effects on the economy.

[24] Two exceptions are that (1) Ebenstein (2012) finds that China's surface water pollution has caused an increase in deaths from digestive cancers; and (2) He and Perloff (2016) find that a deterioration in surface water quality from Water Quality Grade Level I to Level III is associated with higher infant mortality.

Table 1.1 Covariate Balance Between Upstream Downstream Firms

	Mean		Mean Difference
	Downstream	Upstream	≤10km
	(1)	(2)	(4)
Panel A. Time-Invariant Factors			
Year of Opening	1982.55	1984.16	-0.76
	(76.06)	(53.87)	(1.19)
SOE	0.24	0.27	0.02
(1=Yes, 0=Others)	(0.43)	(0.44)	(0.02)
Foreign	0.16	0.19	-0.00
(1=Yes, 0=Others)	(0.37)	(0.39)	(0.02)
Panel B. Pre-2003 Firm-Level Characteristics			
Value Added	8.42	8.48	0.10
(log)	(1.24)	(1.32)	(0.09)
Profit	1.19	2.13	0.30
(billion yuan)	(6.65)	(11.75)	(0.68)
# of Employee	5.06	5.15	0.10
(log)	(1.09)	(1.15)	(0.09)
Capital Stock 1#	8.96	9.04	0.17
(log)	(1.71)	(1.83)	(0.12)
Capital Stock 2#	9.00	9.07	0.15
(log)	(1.70)	(1.83)	(0.12)
Intermediate Input	9.38	9.48	0.11
(log)	(1.32)	(1.38)	(0.10)
Tax	2.22	2.26	0.11
(log)	(2.18)	(2.27)	(0.15)
Obs.	2,190	2,659	

Notes: The sample consists in ASIF firms from 2000–2002 (pre-regulation), that lie within 10km of a water monitoring station. Columns 1–2 report the means and standard deviations of firm characteristics. Column 3 tests the covariate balance between upstream and downstream firms. The difference coefficients are obtained by running OLS regressions of firm characteristics on an upstream dummy and water quality monitoring station and industry fixed effects. Standard errors reported in the parentheses are clustered at the water monitoring station level. # Capital Stock 1 measures the value of fixed assets at the end of year and Capital Stock 2 measures the annual average value of fixed assets.

Table 1.2 RD Estimates of the Impact of Water Quality Monitoring on TFP

	Polluting Industries			Non-Polluting Industries		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Water Quality Monitoring and TFP						
TFP (log) - Polluting Industries	0.36	0.38	0.43	-0.00	0.02	-0.05
	(0.23)	(0.24)	(0.28)	(0.14)	(0.15)	(0.14)
Bandwidth (km)	4.18	3.88	2.88	4.71	4.14	4.19
Panel B: Water Quality Monitoring and Residual TFP						
TFP (log) - Polluting Industries	0.25*	0.25**	0.33**	-0.01	0.00	0.02
(Station FE Absorbed)	(0.14)	(0.13)	(0.15)	(0.09)	(0.10)	(0.11)
Bandwidth (km)	5.80	5.98	4.82	6.02	5.48	4.26
Panel C: Water Quality Monitoring and Residual TFP						
TFP (log) - Polluting Industries	0.31**	0.31**	0.35**	0.02	0.03	0.03
(Station and Industry FE Absorbed)	(0.15)	(0.15)	(0.16)	(0.08)	(0.08)	(0.09)
Bandwidth (km)	6.56	6.54	5.41	5.553	4.918	4.329
Obs.	6,582	6,582	6,582	12,422	12,422	12,422
Kernel	Triangle	Epanech.	Uniform	Triangle	Epanech.	Uniform

Notes: Each cell in the table represents a separate regression. TFP is estimated using Olley and Pakes (1996) method. The discontinuities at monitoring stations are estimated using local linear regressions and MSE-optimal bandwidth proposed by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table 1.3 Heterogeneous Impacts of the Impact of Water Quality Monitoring on TFP

	Residual TFP – Polluting Industries			Residual TFP – Non-Polluting Industries		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: By Ownership						
Private Firms	0.34**	0.37**	0.31*	0.04	0.04	0.03
	(0.17)	(0.18)	(0.18)	(0.08)	(0.08)	(0.09)
Obs.	5,636	5,636	5,636	10,084	10,084	10,084
BW	5.965	5.590	5.087	6.052	6.059	5.537
SOEs	-0.31	-0.16	0.23	-0.13	-0.10	-0.01
	(0.52)	(0.54)	(0.50)	(0.25)	(0.25)	(0.27)
Obs.	635	635	635	1,357	1,357	1,357
BW	4.282	4.474	4.407	4.724	4.545	3.955
Foreign Firms	-0.06	-0.07	-0.11	-0.12	-0.15	0.02
	(0.27)	(0.28)	(0.31)	(0.40)	(0.42)	(0.25)
Obs.	1,104	1,104	1,104	2,427	2,427	2,427
BW	6.927	6.541	5.479	3.287	3.070	4.286
Panel B: By Size						
Small Firm	0.16	0.16	0.14	-0.13	-0.12	-0.14
(Empl<50)	(0.29)	(0.24)	(0.25)	(0.15)	(0.14)	(0.13)
Obs.	1,998	1,998	1,998	4,357	4,357	4,357
BW	7.400	8.096	6.273	3.823	3.985	3.758
Large Firm	0.41***	0.42***	0.40**	-0.01	0.01	0.04
(Empl≥50)	(0.14)	(0.15)	(0.17)	(0.09)	(0.10)	(0.11)
Obs.	5,369	5,369	5,369	9,691	9,691	9,691
BW	4.825	4.738	4.520	4.610	4.674	4.513
Panel C: By Firm Age						
Old Firms	0.33*	0.39**	0.45**	0.05	0.05	0.04
	(0.17)	(0.19)	(0.21)	(0.09)	(0.09)	(0.09)
Obs.	4,481	4,481	4,481	8,373	8,373	8,373
BW	6.695	5.881	4.624	5.432	5.199	4.526
Young Firms	0.48**	0.51**	0.39	-0.03	-0.00	0.07
	(0.19)	(0.21)	(0.26)	(0.16)	(0.18)	(0.20)
Obs.	1,438	1,438	1,438	2,627	2,627	2,627
BW	3.768	3.537	3.798	5.768	5.084	4.357
Kernel	Triangle	Epanech.	Uniform	Triangle	Epanech.	Uniform

Notes: Each cell in the table represents a separate regression. Monitoring station and industry fixed effects are absorbed before estimating regression discontinuity. In columns 1–3, we report the estimated discontinuity for polluting industries, and in columns 4–6, we report the estimated discontinuity for non-polluting industries. Local linear regression and MSE-optimal bandwidth proposed by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods are used for the estimation. Conventional local linear regression discontinuity standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table 1.4. RD Estimates of the Impact of Water Quality Monitoring on TFP by Year

	Residual TFP – Polluting Industries			Residual TFP – Non-Polluting Industries		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Before and After 2003						
Before 2003	0.09	0.10	0.11	0.01	0.01	0.06
	(0.19)	(0.20)	(0.24)	(0.12)	(0.13)	(0.15)
Obs.	2,570	2,570	2,570	4,565	4,565	4,565
After 2003	0.36**	0.35**	0.40**	0.03	0.04	0.07
	(0.16)	(0.16)	(0.17)	(0.08)	(0.09)	(0.10)
Obs.	5,916	5,916	5,916	10,992	10,992	10,992
Panel B. by Year						
Year 2000	-0.09	-0.04	-0.00	-0.22	-0.21	-0.11
	(0.34)	(0.32)	(0.36)	(0.17)	(0.18)	(0.16)
Obs.	1,411	1,411	1,411	2,428	2,428	2,428
Year 2001	-0.02	-0.01	-0.04	-0.07	-0.05	-0.19
	(0.21)	(0.21)	(0.24)	(0.17)	(0.18)	(0.17)
Obs.	1,411	1,411	1,411	2,428	2,428	2,428
Year 2002	0.04	0.09	0.05	0.03	0.01	-0.02
	(0.20)	(0.20)	(0.25)	(0.13)	(0.13)	(0.12)
Obs.	2,106	2,106	2,106	3,644	3,644	3,644
Year 2003	0.30	0.34	0.37*	0.04	0.04	0.04
	(0.29)	(0.29)	(0.21)	(0.16)	(0.16)	(0.15)
Obs.	2,367	2,367	2,367	3,888	3,888	3,888
Year 2004	0.12	0.14	0.21	0.08	0.06	0.06
	(0.30)	(0.32)	(0.31)	(0.11)	(0.11)	(0.11)
Obs.	3,288	3,288	3,288	5,509	5,509	5,509
Year 2005	0.31	0.35	0.35	-0.04	-0.05	-0.06
	(0.24)	(0.25)	(0.26)	(0.15)	(0.15)	(0.15)
Obs.	3,750	3,750	3,750	6,296	6,296	6,296
Year 2006	0.48**	0.52**	0.61**	0.01	0.01	0.03
	(0.22)	(0.25)	(0.27)	(0.14)	(0.15)	(0.16)
Obs.	3,981	3,981	3,981	6,969	6,969	6,969
Year 2007	0.37**	0.38*	0.42*	0.14	0.15	0.17*
	(0.19)	(0.20)	(0.22)	(0.09)	(0.09)	(0.10)
Obs.	4,460	4,460	4,460	8,103	8,103	8,103
Kernel	Triangle	Epanech.	Uniform	Triangle	Epanech.	Uniform

Notes: Each cell in the table represents a separate regression. Monitoring station and industry fixed effects are absorbed before estimating regression discontinuity. In columns 1-3, we report the estimated discontinuity for polluting industries, and in columns 4-6, we report the estimated discontinuity for non-polluting industries. Local linear regression and MSE-optimal bandwidth proposed by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods are used for the estimation. Conventional local linear regression discontinuity standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table 1.5 Instrumental Variable Estimation using Hydrological Stations

	Polluting Industries		Non–Polluting Industries	
	Upstream	TFP (log)	Upstream	TFP (log)
	(1)	(2)	(3)	(4)
Upstream Hydrological Station	0.38**		0.30**	
	(0.18)		(0.14)	
Upstream Monitoring Station		-0.35**		-0.08
		(0.15)		(0.16)
Specification	1st Stage	2SLS	1st Stage	2SLS
Station FE	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y
Observations	4,445	4,462	8,976	8,981
F Statistic	10.48	0.03	22.82	1.18

Notes: Each column in the table represents a separate regression. We define "upstream monitoring station" as a dummy indicator for whether a firm is upstream from a monitoring station within a 10 km range, and similarly, we define "upstream hydrological station" as a dummy indicator for whether a firm is upstream from a hydrological station within a 10 km range. Our outcome of interest is firm-level TFP estimated using Olley and Pakes (1996) method, our endogenous variable is "upstream monitoring station", and our instrumental variable is "upstream hydrological station". We present first-stage results and IV 2SLS results separately for firms in polluting industries (columns 1 and 2) and firms in non-polluting industries (columns 3 and 4). Monitoring station fixed effects are controlled for in all specifications. Standard errors are clustered at the monitoring station level. * significant at 10% ** significant at 5% *** significant at 1%.

Table 1.6 RD Estimates using Placebo Downstream Firms

	Polluting Industries			Non-Polluting Industries		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Water Quality Monitoring and TFP						
TFP (log) - Polluting Industries	0.36	0.44*	0.26	-0.18	-0.20	-0.13
	(0.23)	(0.26)	(0.29)	(0.16)	(0.15)	(0.18)
Panel B: Water Quality Monitoring and Residual TFP						
TFP (log) - Polluting Industries	0.48**	0.52**	0.61***	0.13	0.11	0.14
(Station and Industry FE Absorbed)	(0.20)	(0.21)	(0.23)	(0.13)	(0.11)	(0.12)
Obs.	4,435	4,435	4,435	8,001	8,001	8,001
Kernel	Triangle	Epanech.	Uniform	Triangle	Epanech.	Uniform

Notes: Each cell in the table represents a separate regression, where each control firm is replaced by its best match in the whole sample from a pre-2003 nearest neighbour matching (based on TFP, industry, and other basic characteristics). TFP is estimated using Olley and Pakes (1996) method. The discontinuities at monitoring stations are estimated using local linear regression and MSE-optimal bandwidth proposed by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the coefficients in columns 1–5 and conventional local linear regression discontinuity standard errors clustered at the monitoring station level are reported in columns 6-8. * significant at 10% ** significant at 5% *** significant at 1%.

Table 1.7 Robustness Checks: Impact of Water Quality Monitoring on TFP

	TFP – Polluting Industries		
	(1)	(2)	(3)
Panel A. Alternative Ways to Estimate RD and Standard Errors			
Bias-corrected RD Estimates	0.35**	0.34**	0.38**
	(0.15)	(0.15)	(0.16)
Bias-corrected Robust Estimates	0.35*	0.34*	0.38**
	(0.19)	(0.19)	(0.19)
Panel B. Alternative Ways to Choose Optimal Bandwidth			
Bandwidth Chosen by MSE-Two Selector	0.30**	0.29*	0.25
	(0.15)	(0.15)	(0.17)
Bandwidth Chosen by MSE-Sum Selector	0.31**	0.30**	0.34**
	(0.15)	(0.15)	(0.16)
Bandwidth Chosen by CER-D Selector	0.38**	0.40**	0.43**
	(0.19)	(0.19)	(0.20)
Bandwidth Chosen by CER-Two Selector	0.35**	0.39**	0.48**
	(0.17)	(0.17)	(0.20)
Bandwidth Chosen by CER-Sum Selector	0.37**	0.39**	0.44**
	(0.18)	(0.19)	(0.20)
Panel C. Placebo Tests			
Move Monitoring Stations Upstream by 5km	0.12	0.13	0.11
	(0.16)	(0.16)	(0.16)
Move Monitoring Stations Upstream by 10km	-0.08	-0.09	-0.08
	(0.11)	(0.11)	(0.12)
Move Monitoring Stations Downstream by 5km	0.13	0.15	0.11
	(0.09)	(0.09)	(0.11)
Move Monitoring Stations Downstream by 10km	0.03	0.05	0.07
	(0.16)	(0.15)	(0.17)
Kernel	Triangle	Epanech.	Uniform

Notes: Each cell in the table represents a separate regression. Monitoring station and industry fixed effects are absorbed before estimating regression discontinuity. Conventional local linear regression discontinuity standard errors clustered at the monitoring station level are reported below the estimates (except Panel A). Local linear regression and MSE-optimal bandwidth selected by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods are used for the estimation (except Panel B). Monitoring station and industry fixed effects are absorbed before estimating regression discontinuity. * significant at 10% ** significant at 5% *** significant at 1%.

Table 1.8 Channels: RD Estimates on other Measures

	Conventional Local RD		Bias-Corrected RD		Bias-Corrected Robust	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Output Related						
Value-Added	0.10	0.09	0.04	0.04	0.04	0.04
(log)	(0.20)	(0.21)	(0.20)	(0.21)	(0.26)	(0.27)
Profit	1.07	1.18	1.34*	1.49*	1.34	1.49
(10 million yuan)	(0.74)	(0.80)	(0.74)	(0.80)	(0.89)	(0.95)
Panel B. Input Related						
Employees	-0.11	-0.11	-0.16	-0.13	-0.16	-0.13
(log)	(0.15)	(0.14)	(0.15)	(0.14)	(0.19)	(0.19)
Intermediate Input	-0.07	-0.04	-0.14	-0.11	-0.14	-0.11
(log)	(0.21)	(0.23)	(0.21)	(0.23)	(0.29)	(0.30)
Capital Stock 1 #	-0.44	-0.42	-0.55**	-0.54**	-0.55	-0.54
(log)	(0.27)	(0.27)	(0.27)	(0.27)	(0.34)	(0.34)
Capital Stock 2 #	-0.50*	-0.48*	-0.63**	-0.61**	-0.63*	-0.61*
(log)	(0.28)	(0.28)	(0.28)	(0.28)	(0.36)	(0.35)
Panel C. Naïve Labor and Capital Productivity						
VA/Employee	0.23	0.31	0.26	0.36	0.26	0.36
(log)	(0.20)	(0.22)	(0.20)	(0.22)	(0.25)	(0.27)
VA/Cap Stock 1	0.27***	0.28***	0.26**	0.29***	0.26**	0.29**
(log)	(0.10)	(0.11)	(0.10)	(0.11)	(0.12)	(0.13)
VA/Cap Stock 2	0.23**	0.26**	0.23**	0.27**	0.23*	0.27**
(log)	(0.11)	(0.11)	(0.11)	(0.11)	(0.14)	(0.14)
Panel D. Porter Hypothesis						
R&D	-0.09	-0.10	-0.07	-0.10	-0.07	-0.10
(log)	(0.35)	(0.37)	(0.35)	(0.37)	(0.55)	(0.57)
Kernel	Triangle	Epanech.	Triangle	Epanech.	Triangle	Epanech.

Notes: Each cell in the table represents a separate regression. Monitoring station and industry fixed effects are absorbed before estimating regression discontinuity. We use post-2003 collapsed data to estimate the regression discontinuities. Local linear regression and MSE-optimal bandwidth proposed by Calnoico et al. (2014) and Calonico (2017) for different kernel weighting methods are used for the estimation. Station and industry fixed effects are controlled in all regressions. Standard errors are clustered at the monitoring station level, and reported below the estimates. # Capital Stock 1 is measured by the value of fixed assets at the end of each year and Capital Stock 2 is the annual average value of fixed assets. * significant at 10% ** significant at 5% *** significant at 1%.

Table 1.9 Political Economy of Water Quality Monitoring

	Conventional Local RD		Bias-Corrected RD		Bias-Corrected Robust	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. "Double Standard"						
Waste Discharge Fee	-1.15**	-1.07**	-1.41***	-1.32**	-1.41**	-1.32**
(log)	(0.51)	(0.53)	(0.51)	(0.53)	(0.57)	(0.60)
Panel B. Strong vs. Weak Political Incentives						
TFP (log) - Strong Incentive	0.57***	0.59***	0.63***	0.66***	0.63***	0.66***
	(0.19)	(0.20)	(0.19)	(0.20)	(0.21)	(0.23)
TFP (log) - Weak Incentive	0.01	0.08	0.00	0.07	0.00	0.07
	(0.23)	(0.24)	(0.23)	(0.24)	(0.29)	(0.31)
Panel C. Automatic vs. Manual Monitoring Stations						
TFP (log) - Automatic Stations	0.92	1.01*	1.11*	1.22**	1.11	1.22*
	(0.59)	(0.57)	(0.59)	(0.57)	(0.74)	(0.71)
TFP (log) - Manual Stations	0.26*	0.26*	0.27*	0.27*	0.27	0.27
	(0.15)	(0.15)	(0.15)	(0.15)	(0.18)	(0.18)
Kernel	Triangle	Epanech.	Triangle	Epanech.	Triangle	Epanech.

Notes: Each cell in the table represents a separate regression. Monitoring station and industry fixed effects are absorbed before estimating regression discontinuity. We focus on polluting firms and use post-2003 collapsed data to estimate the regression discontinuities. Local linear regression and MSE-optimal bandwidth proposed by Calnoico et al. (2014) and Calonico (2017) for different kernel weighting methods are used for the estimation. Panel A examines how tax and waste discharge fee collected by the government differ between upstream and downstream firms. Panel B estimates the discontinuities separately using the subsamples where the Prefecture Party Secretary has or does not have strong promotion incentives (age<=57 vs. age >57). Panel C estimates the discontinuities separately for automatic and manual monitoring stations. * significant at 10% ** significant at 5% *** significant at 1%.

Table 1.10 RD Estimates of the Impact of Water Quality Monitoring on Emissions

	Conventional Local RD		Bias-Corrected		Bias-Corrected Robust	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: COD Emission						
COD Emission (log)	0.83*	0.75*	0.99**	0.92**	0.99**	0.92*
	(0.44)	(0.42)	(0.44)	(0.42)	(0.49)	(0.47)
COD Emission Intensity (log)	0.55**	0.49*	0.68**	0.62**	0.68**	0.62**
	(0.27)	(0.26)	(0.27)	(0.26)	(0.32)	(0.31)
Panel B: Wastewater Discharge						
Waste Water Discharge (log)	0.39	0.39	0.49	0.50	0.49	0.50
	(0.33)	(0.35)	(0.33)	(0.35)	(0.40)	(0.42)
Waste Water Discharge Intensity (log)	0.34*	0.33*	0.42**	0.41**	0.42*	0.41*
	(0.20)	(0.20)	(0.20)	(0.20)	(0.23)	(0.22)
Bandwidth Selector	MSE	MSE	MSE	MSE	MSE	MSE
Obs.	9,888	9,888	9,888	9,888	9,888	9,888
Kernel	Triangle	Epanech.	Triangle	Epanech.	Triangle	Epanech.

Notes: Each cell in the table represents a separate regression. Monitoring station fixed effects are absorbed before estimating regression discontinuity. Local linear regression and MSE-optimal bandwidth selected by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods are used for the estimation. Conventional local linear regression discontinuity standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table 1.11 Predicted Effects of Water Quality Monitoring on TFP and COD

	Model 1	Model 2	Model 3
	(1)	(2)	(3)
Panel A. TFP Effects for Large and Small Firms (Measured by Industry Output Value)			
20% (log Rev ~ 9.01)	0.14	0.10	0.11
	(0.24)	(0.23)	(0.23)
40% (log Rev ~ 9.58)	0.27	0.32	0.32
	(0.23)	(0.22)	-0.22
60% (log Rev ~ 10.16)	0.42*	0.50**	0.49**
	(0.22)	(0.21)	0.21
80% (log Rev ~ 10.92)	0.62***	0.66***	0.65***
	(0.22)	(0.20)	(0.20)
Panel B. COD Effect for Large and Small emitters (Measured by COD Emissions)			
20% (log COD ~ 5.97)	0.14	0.12	0.19
	0.44	(0.43)	(0.44)
40% (log COD ~ 7.46)	0.93**	0.96**	0.99**
	(0.43)	(0.43)	(0.44)
60% (log COD ~ 8.70)	1.59***	1.62***	1.59***
	(0.43)	(0.43)	(0.43)
80% (log COD ~ 10.18)	2.38***	2.34***	2.27***
	(0.43)	(0.43)	(0.42)
Heterogeneity Specification	Linear	Quadratic	Cubic

Notes: This table reports the predicted effects of water quality monitoring on TFP and COD emission intensity. Monitoring station and industry fixed effects are absorbed before estimating regression discontinuity. In Panel A, we explore the TFP heterogeneity at different revenue levels, and in Panel B, we explore the COD intensity heterogeneity at different COD emission levels. We use parametric RD to estimate the heterogeneous effects using different heterogeneity functional forms. We choose the polynomial RD specifications that generate the closest estimates to the non-parametric estimates reported in Table 2 and Table 8 as the baselines. We then include interactions to test the heterogeneity. "Linear" means that we use a linear interaction between the downstream dummy and log revenue (or log COD), and "quadratic" means we interact the downstream dummy with a quadratic function of log revenue (or log COD). Panel A shows that the monitoring effect is only significant for large firms, and Panel B shows that the monitoring effect is only significant for large emitters.

Table 1.12 Economic Costs of COD Abatement

	Conventional		Bias-Corrected	
	(1)	(2)	(3)	(4)
Panel A. Estimated Effect of Water Quality Monitoring				
Effect on log TFP	0.31**	0.31**	0.35**	0.34**
	(0.15)	(0.15)	(0.15)	(0.15)
Effect on log COD Emission	0.83*	0.75*	0.99**	0.92**
	(0.44)	(0.42)	(0.44)	(0.42)
Effect on log COD Emission Intensity	0.55**	0.49*	0.68**	0.62**
	(0.27)	(0.26)	(0.27)	(0.26)
Panel B. Estimated Economic Costs Estimates:				
TFP Loss if all Polluting Firms are Monitored	26.66%	26.66%	29.53%	28.82 %
TFP Loss per 10% COD Emission Abatement	2.49%	2.75%	2.35%	2.46%
TFP Loss per 10% COD Emission Intensity Reduction	3.75%	4.21%	3.43%	3.65%
Total Output Loss if all Polluting Firms are Monitored (billion CNY)	3988.9	3988.9	4599.6	4444.6
Total Output Loss in the Polluting Industry during the 11th Five-Year Plan (billion CNY), A	351.98	390.86	332.60	348.16
Total Output Loss in the Polluting Industry per 2.5% COD Abatement (billion CNY), A	68.64	76.01	64.95	67.91
Total Output Loss in the Polluting Industry per 10% COD Abatement (billion CNY), A	279.79	310.48	264.48	276.77
Total Output Loss in the Polluting Industry per 2.5% COD Abatement (billion CNY), B	242.91	269.00	229.85	240.34
Total Output Loss in the Polluting Industry per 10% COD Abatement (billion CNY), B	990.2	1098.8	936.0	979.5
Kernel	Tri	Epa	Tri	Epa
Gross Output Value in the Polluting Industry in 2006 (billion CNY), A	10975.7			
Gross Output Value in the Polluting Industry in 2015 (billion CNY), B	38844.9			

Notes: The gross output values were obtained from the website of the National Bureau of Statistics. A: calculation is based on gross output value (of industries above designated size) in 2006; B: calculation is based on gross output value (of industries above designated size) in 2015.

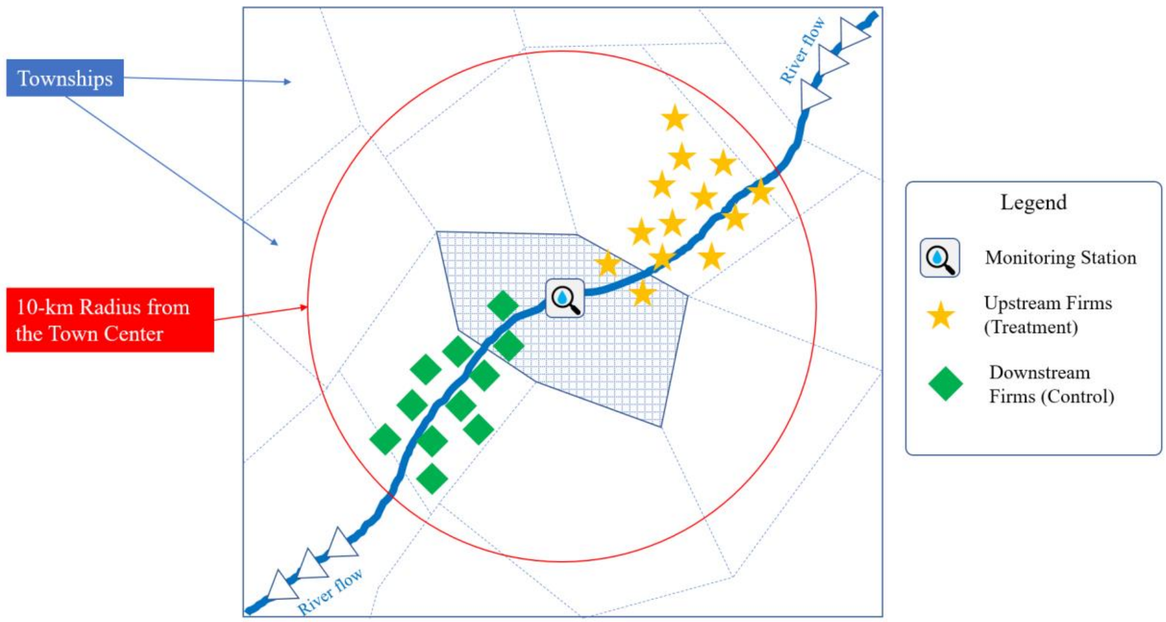


Figure 1.1 Illustrating the Identification Strategy

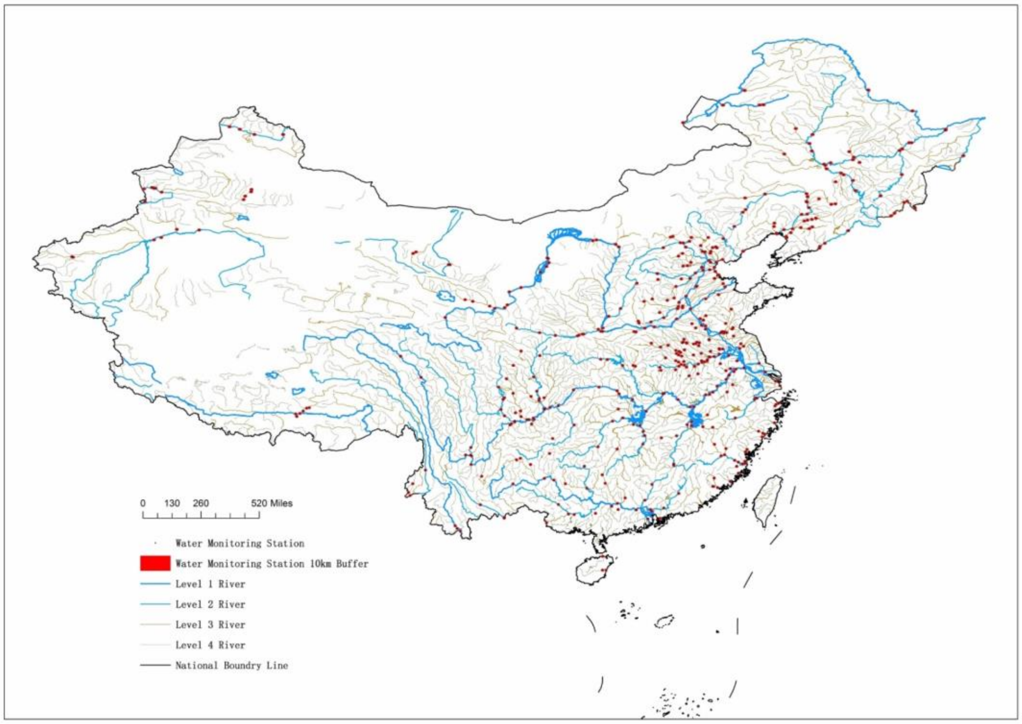
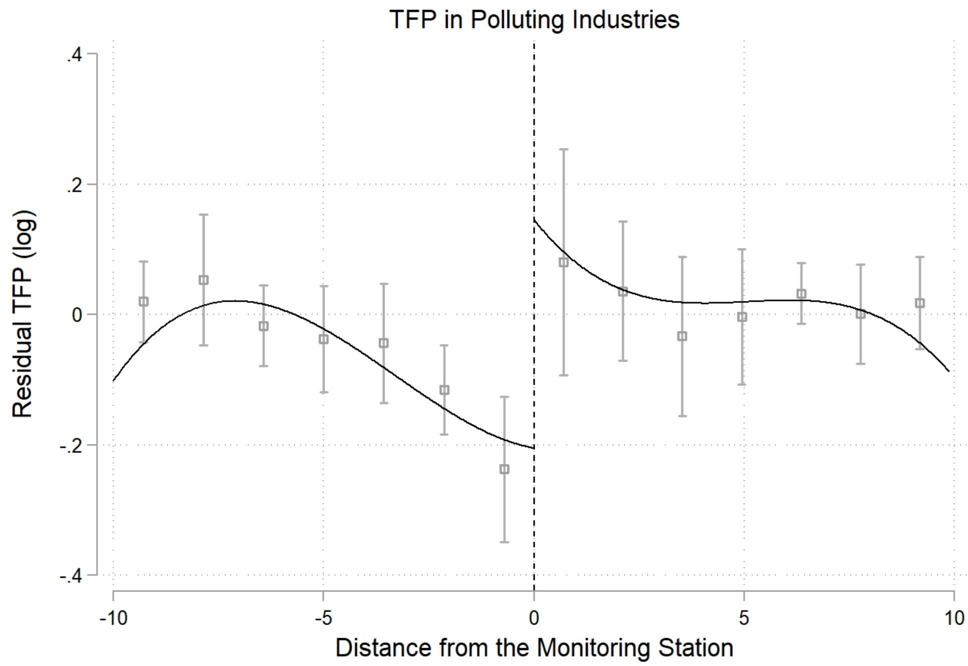


Figure 1.2 Distribution of Surface Water Quality Monitoring Stations

Panel A: TFP in Polluting Industries



Panel B: TFP in Non-Polluting Industries

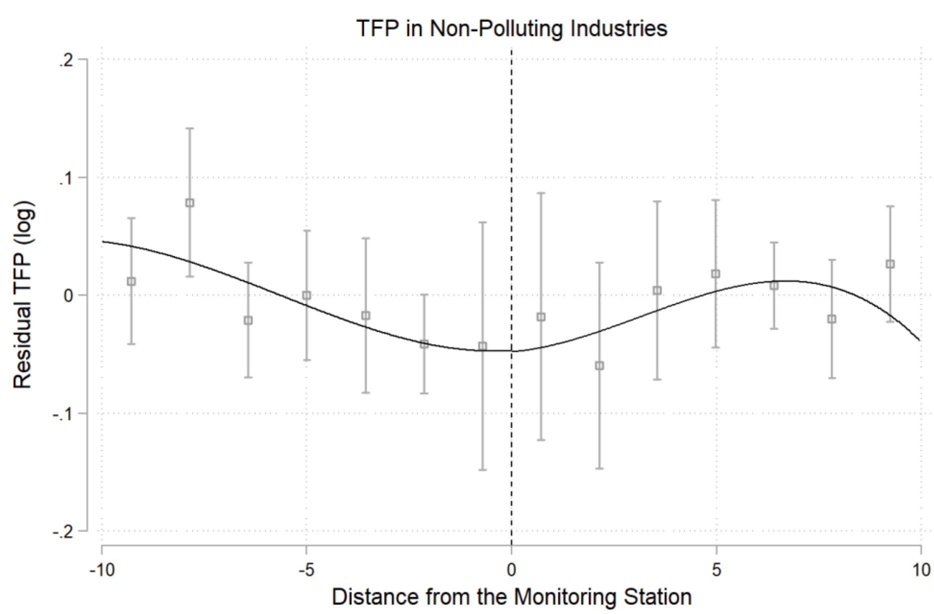


Figure 1.3 RD Plot: Effects of Water Quality Monitoring on TFP

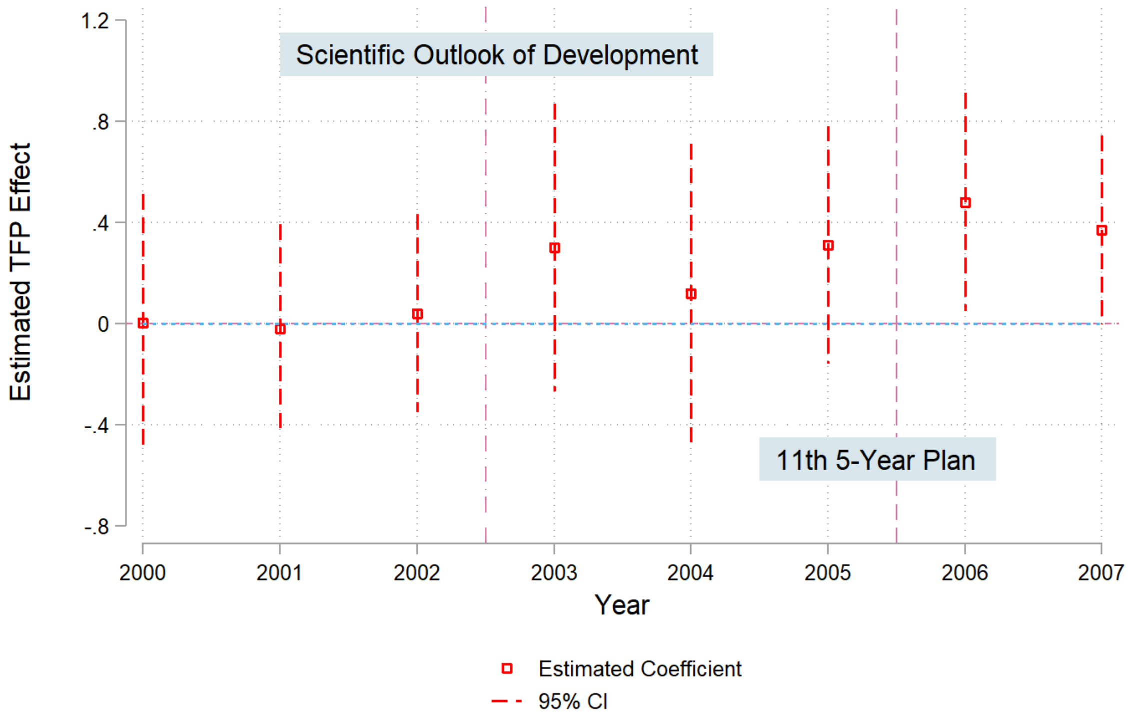


Figure 1.4 RD Estimates by Year

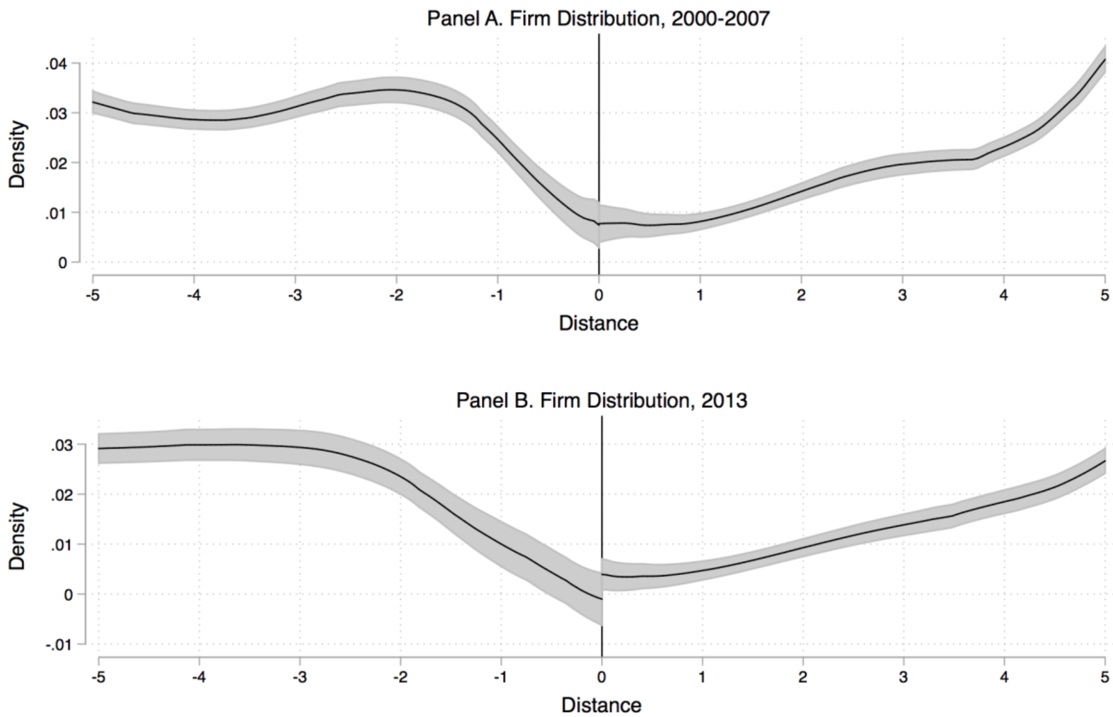


Figure 1.5 Distribution of Firms

2 Fiscal Competition and Coordination: Evidence from China

2.1 Introduction

In his seminal 1956 paper, Charles Tiebout proposed the idea of "Tiebout Sorting," under which different communities provide different bundles of public goods, and mobile households sort to the communities whose bundles best satisfy their needs. Under the Tiebout framework, inter-jurisdictional fiscal competition leads to an efficient provision of public goods and an optimal allocation of households across different communities. The idea of Tiebout sorting is not unique to labor, and has been generalized to the competition for other production factors, such as mobile capital or the locational choices of firms (White, 1976; Wellisch, 1996). Not surprisingly, the Tiebout results have traditionally been used to support fiscal federalism.

In stark contrast, a large literature on fiscal competition argues the exact opposite: competition among local jurisdictions distort the provision of public goods, or the allocation of production factors, or both. The key insight behind this argument is the existence of fiscal externalities: a local jurisdiction regards the outflow of its production factors as a pure cost, but fails to internalize the potential benefits to other jurisdictions receiving extra production factors due to such relocations. This literature started with tax competition, where competing jurisdictions fail to internalize the positive externalities of their potential capital outflow, therefore tending to set an inefficiently low tax rate (Oates, 1970), and has been generalized to various contexts (Wildasin, 1991; Wilson and Gordon, 2003; Saavedra, 2000). Typically, this literature tends to focus on the inefficiencies associated with inter-jurisdictional competition, and thus advocate for a centralized coordinating policy to internalize the externalities created by resource flows across local jurisdictions.

How empirically relevant is fiscal competition? Would fiscal competition lead to a higher or lower level of public good provision, as compared to fiscal coordination? How different is the allocation of production factors under these two regimes? Rigorous answers to these questions are crucial for the evaluation of different theoretical frameworks, and even more importantly, central to the long-standing debates on fiscal federalism.

Identifying causal evidence on the existence and implications of fiscal competition and coordination is difficult, mainly due to the lack of a good counterfactual: jurisdictions that compete and those that coordinate are usually also different in many other unobservable ways. Therefore, despite rich suggestive evidence on the existence and implications of these two opposite regimes, a credible causal interpretation is still largely missing in the literature.

In this paper, we intend to contribute to the empirical studies on fiscal competition, by providing causal evidence on the existence and implications of fiscal competition and coordination, using a unique empirical setting created by two novel national programs in China.

The first national program is the massive construction of "Special Economic Zones (SEZs)," under which certain rural counties are selected by prefecture city governments, and given policy priorities to quickly become urbanized and industrialized. China built more than 1300 SEZs since the 1980s, and within each prefecture city (henceforth, city),[25] there exists at least one SEZ. Past studies have established that the SEZs have significantly higher labor productivity than their neighboring agricultural counties (Wang, 2013; Cheng 2015), and the successful construction of a SEZ is highly rewarding for the prefecture city it belongs to, both politically and economically.

The second national program is the "Rural School Consolidation (RSC)" movement in China, under which the prefecture city governments were encouraged to rearrange the distribution of primary schools within their constituency, by closing small primary schools and consolidating them into larger schools, to realize economy of scale in education. Within one decade, under the RSC program, China closed more than 50% of its rural primary schools, which led to severe social problems and heated debates. Generally believed to have been over-implemented, the RSC program was terminated by the central government in 2012.

The combination of these two national programs creates an ideal empirical setting to identify the causal effects of fiscal competition and coordination, as illustrated in Figure 2.1. In this setting, the SEZ lies on the border of prefecture cities A and B, and is more productive than its neighboring counties. The SEZ belongs to city A, and in its neighborhood, some counties (treated) belong to the same prefecture city, A; while other counties (control) belong to a different prefecture city, B.

Before the RSC program started, the provision of public goods was highly decentralized at the county level, so the treated and control counties each made their own decisions on local public goods, and neither of them would internalize the positive externalities of their labor flowing to the (more productive) SEZ. Therefore, the treated counties, the control counties, and the SEZ engaged in fiscal competition over local public good provision, inducing the reallocation of mobile labor. Because the treated and control counties are spatially adjacent, and they make the same competitive decision (provide high levels of public good to prevent labor from leaving for the SEZ), we expect that before the RSC started, the treated and control counties were well balanced in every measure, including the levels of public good provision and population outflow.

After the RSC program started, the provision of primary schools became centralized at the prefecture city level. Since prefecture city A owns the SEZ, it internalizes the positive externalities of migration flows from the treated counties to the SEZ. Therefore, the leader of prefecture city A now makes a joint decision on primary schools for both the treated counties and the SEZ, and he has the incentives to coordinate them: by providing fewer primary schools in the treated counties, more labor would concentrate to the SEZ, increasing aggregate production for prefecture city A. In contrast, since prefecture city B does not own the SEZ, it does not internalize the positive externalities of migration flows from the control counties to the SEZ. Therefore, the leader of prefecture city B makes a decision on primary schools only for the control counties, and he still has incentive to compete with the SEZ: by providing more primary schools in the control counties, fewer workers would leave city B for the SEZ, increasing production for prefecture city B. Due to the opposite incentives of prefecture cities A and B, after the RSC program started, there should emerge a discontinuous gap between the spatially adjacent treated and control counties: the treated counties should close more schools and lose more population, as compared to the control counties.

Since the RSC program only applies to primary schools but not to other types of public good, even after the RSC started, the provision of other types of local public good remains decentralized at the county

[25] In the Chinese system, prefecture city is the level of government between province (comparable to state in the U.S. system) and county (comparable to county in the U.S. system).

level. Since both the treated and control counties still do not internalize the benefits of the SEZ, they keep on making competitive decisions for those other types of public good. While the differential amounts of primary schools imposed by the prefecture city governments might have indirect impacts on the decision for other types of public good, since the county governments have control over many other different types of public good, such indirect impacts would be split and attenuated between all those other types of public good. Therefore, for each specific type of local public good (other than primary schools), the impact of the RSC program is expected to be minimal, which means that other types of public good would likely remain balanced between the treated and control counties.

In addition, given that economic growth is a key factor in the Chinese meritocracy system, politicians with stronger promotion incentives will likely care more about local GDP. Therefore, during the post-RSC period, if the leader of prefecture city A has strong promotion incentives, we would expect to see a higher level of coordination between the treated counties and the SEZ, which means a higher level of primary school closure in the treated counties; if the leader of prefecture city B has strong promotion incentives, we would expect to see a higher level of competition between the control counties and the SEZ, which means a lower level of primary school closure in the control counties.

To formalize the aforementioned intuitions, we present a simple model of asymmetric fiscal competition, which mimics our empirical setting, and transforms the intuitions into four key testable hypotheses: (1) before the RSC started, the treated and control counties should be well balanced in every measure; (2) after the RSC started, the treated counties closed more primary schools and had more population outflow than the control counties; (3) after the RSC started, except for primary schools, other types of local public good remain balanced between the treated and control counties; (4) after the RSC started, if the leader of prefecture city A has strong promotion incentives, the treated counties would close even more primary schools; if the leader of prefecture city B has strong promotion incentives, the control counties would close even fewer primary schools.

Using a unique panel dataset of Chinese townships,[26] we apply a spatial discontinuity design as illustrated in Figure 2.1, and find strong empirical evidence consistently confirming all four testable hypotheses. In addition, we also adopt various alternative specifications and placebo tests to check for robustness, explore for heterogeneity, and rule out competing explanations.

This paper relates to several strands of literature.

First and foremost, it contributes to the large literature on fiscal competition and strategic interactions among local governments, which falls more generally into the literature on decentralization and fiscal federalism. Tiebout (1956) first argues that mobile households can lead to regional competition that improves welfare, and this argument was later extended to mobile capital and firms (White, 1976; Wellisch, 1996). Using a case of tax competition, Oates (1972) shows that inter-governmental competition can lead to distortions in the public sector (tax rate in his case) and misallocation of the production factor (capital in his case). This idea was later formalized (Zodrow and Mieszkowski, 1986), and extended to many other forms of competition, including income redistribution (Wildasin, 1991), government expenditure (Wilson and Gordon, 2003), environmental policies (Fredriksson and Millimet, 2002), and welfare transfers (Saavedra, 2000). It has also been argued that regional competition may increase the provision of public good by "taming Leviathan governments" (Brennan and Buchanan, 1980), which is consistent with various stylized facts pointed out in the literature (Oates, 1985; Oates, 1989). While there

[26] The county dataset does not contain information on primary schools, so our unit of analysis is at the township level, which in the Chinese system is a smaller unit below county.

exists a large body of theoretical literature with different forms and implications, solid causal evidence directly testing these models are surprisingly inadequate. As pointed out by Brueckner (2003), existing empirical literature typically justifies the existence of regional competition by testing some indirect implications of the theoretical models, and usually suffers from serious endogeneity problems. Our paper intends to estimate the causal effects of fiscal competition and coordination on public good provision and labor allocation. We address the identification threats mentioned in Bruckner (2003) using a spatial discontinuity design, and find that when governments are Leviathan, competition would increase public good supply and decrease the level of migration, which are consistent with the propositions of Brennan and Buchanan (1980).

Second, this paper adds to the empirical tests of the Tiebout Sorting hypothesis. Since Tiebout (1956), there has been continuous efforts trying to justify Tiebout Sorting empirically. Earlier works generally focus on testing some indirect implications of the model, including capitalization and hedonics (Oates, 1969; Rosen 1974), efficient provision of public good (Brueckner, 1982), stratification of demand for public goods and the link between diversity in income and provision of public good (Epple and Sieg, 1999; Rhode and Strumpf, 2003). More recently, the empirical literature tends to directly test the Tiebout mechanism using quasi-experimental changes in public goods, for example, among many others, Urquiola (2005) shows that inter-district school choice leads to sorting, Banzhaf and Walsh (2008) finds that people sort toward communities with increased air quality. In the context of China, a paper closely related to ours is Xing (2014), which uses individual-level census data to show that in villages where the RSC program is implemented (local primary school is closed), parents with school-aged children become much more likely to migrate out and seek for external schooling opportunities for their children. Although not its main focus, our paper also shows that when RSC program is implemented, students, teachers, and parents would sort after schools, which confirms the Tiebout hypothesis in the context of China.

Third, this paper also speaks to the literature documenting China's state-led style of urbanization. The central government of China always had the temptation to boost urbanization in the country (Fan el al., 2012), and it has been estimated that the revenues from converting agricultural land into industrial use could account for 60-80% of the total government revenues in many parts of China (Zhu, 2005). As a result, promoting urbanization becomes highly profitable for the local governments, both politically and economically, and is highly prioritized by the local officials (Han and Kung, 2015). Among many attempts to promote urbanization, constructing SEZs has been proved highly successful in many regions (Wang, 2013), but there are also cases of failure where the SEZs could not concentrate enough population and ended up being "ghost townships" (Chen and Li, 2015). This paper documents that the relocation of public good is an effective method through which many local governments boost urbanization for their SEZs, which is new to this literature, and could help improve our understanding of the patterns and welfare implications of the state-led urbanization in China.

The remaining parts of this paper are organized as follows. Section 2 discusses the empirical context, provides qualitative evidence, and presents the identification strategy. Section 3 presents a simple model of asymmetric fiscal competition, which mimics the empirical setting, and formalizes the intuitions into four main testable hypotheses. Section 4 introduces the three datasets used in the paper. Section 5 formally examines the four testable hypotheses, and presents the main findings of the paper. Section 6 shows additional analysis to check for robustness, explore heterogeneity, and rule out alternative explanations. Section 7 concludes.

2.2 Context

The empirical context of this paper is created by the combination of two national programs in China, the massive construction of the "Special Economic Zones (SEZs)," and the "Rural School Consolidation (RSC)" movement, which we introduce in sections 2.1 and 2.2, respectively. In section 2.3, we summarize anecdotal evidence on the interactions between these two national programs, which provides qualitative support for the existence of fiscal competition and coordination in our context. In section 2.4, we present the spatial-discontinuity design, which is the main identification strategy of this paper.

Special Economic Zones

As a place-based industrial policy tool, various kinds of Special Economic Zones (SEZs) have been constructed in many different countries, and overall, they show a mixed record of success (Farole & Akinci, 2011).

In the case of China, SEZs are usually county-level units based on land retired from agricultural production, which are given tax cuts and other preferential policies by the corresponding prefecture city governments to attract FDI and domestic investments.

The Chinese experience with SEZs has been highly successful: Wang (2013) finds that SEZs bring higher levels of FDI to the corresponding prefecture cities in general, without crowding out domestic investment; Cheng (2014) finds that transforming an agricultural county into a SEZ increases local GDP level by about 6 percent after 5 years; Alder et al. (2015) find that the establishment of a SEZ is associated with an increase in the level of per capita GDP by about 18% for the host prefecture city.

Since the promotion of local officials largely depends on their ability to boost economic growth (Li and Zhou, 2005), there also exists high political returns to the successful construction of a SEZ. Driven by the high economic and political benefits, China constructed more SEZs than any other country in the world (Boyenge, 2007). From 1980 to 2006, more than 1300 SEZs were built, with every prefecture city constructing at least one SEZ.

As shown in the existing literature, due to reasons such as higher physical capital, agglomeration effects, tax cuts, and better technology, the marginal product of labor is significantly higher in the SEZs than in their neighboring counties, for both high-skilled and low-skilled workers (Wang, 2013; Alder et al., 2015).

Rural School Consolidation

The "One-Child Policy" in China led to steady decrease of school-aged children in rural areas since the 1980s, and the massive rural-urban migration further exacerbated this problem. Yet, more rural primary schools were constructed during the 1980s and 1990s under the "Compulsory Education" policy. As a joint result of these three forces, by 2000, there was an obvious surplus of rural primary schools, and some schools could hardly enroll enough students to be sustainable.

To make efficient use of educational resources and realize economy of scale in education, in 2001, the central government issued the *Decisions on the Reform and Development of Basic Education,* which required the prefecture city governments to close small rural primary schools and consolidate them into larger schools. While starting from good intentions, this "Rural School Consolidation (RSC)" program was over-implemented by the local governments: in only one decade, China closed more than 50% of its rural primary schools,[27] which means closing more than 4 schools every hour, and has continued over ten years. Not surprisingly, such ultra-large number of school closure far exceeded the magnitude of the

[27] From 553622 in 2000 to 257410 in 2010.

shortage of rural students, and caused a series of negative aftermath for rural families: increased distance from home to schools, more frequent school bus accidents in rural areas, the number of boarding schools insufficient to meet increased boarding demands, higher schooling costs for rural families, and higher dropout rates, etc (Yang, 2012). These problems have been pointed out since 2005, and the central government warned the local governments twice on not over-implementing the RSC program, first in 2006 and again in 2010, which, however, failed to slow down the striking trend of school closure in the rural areas. As a result, the central government chose to terminate the RSC program in 2012, and specifically emphasized that rural schools should be protected from being closed by local governments.[28]

It is critical to note that before the RSC program started, the provision of rural education, like other types of local public good, was highly decentralized: decisions were made almost exclusively at the county and lower levels, and the prefecture city government had very limited ability to interfere;[29] however, after the introduction of the RSC program, the prefecture city government suddenly obtained the authority to manage the distribution of primary schools across different counties. Therefore, the RSC program could be regarded as introducing a planner to make a joint decision on primary schools for all the counties in the same prefecture city. Meanwhile, since the RSC program only applies to primary schools, other types of public good are not directly influenced by this program.

Qualitative Evidence on the Interactions between SEZ and RSC

There exists rich anecdotal evidence from various sources that documents the interactions between the SEZ program and the RSC program: after the RSC program started, the prefecture city governments obtained the power to rearrange the distribution of primary schools across different counties within their constituency. To induce agricultural labor to migrate to their own SEZs, which have higher marginal products of labor, the prefecture city governments intentionally closed schools in the neighborhoods of their SEZs, and consolidated their resources into the schools in the neighboring SEZs.

In their guidelines on the implementation of the RSC program, nearly all the provinces emphasize the intention of using the RSC program to promote urbanization (for examples, see documents of Hebei Province, Fujian Province, and Henan Province).[30] A number of media reports also confirm that prefecture city governments utilize rural school consolidation to concentrate students and their families into the SEZs to promote urban and industrial development (see examples in Chengdu, Shuyang, and Lishui).[31] In recent evaluations of the RSC program that were published on the state-owned media, researchers argue that

---

[28] For more details, see the State Council (2006, 2010, 2012).

[29] For more discussion on the decentralized provision of public goods in rural China, see Wang (2008), Sato (2008), and Martinez-Vazquez et al. (2008).

[30] Hebei: http://www.fnjy.net/Item/5427.aspx, 04-29-2006

Fujian: http://www.fjshjy.net/zcfg/zcwj/201012/t20101209_74634.htm, 11-22-2010

Henan: http://xsaq.haedu.cn/2010/10/12/1286866728093.html, 12-17-2010

[31] Chengdu: http://www.jyb.cn/basc/ts/201005/t20100513_359799.html, 05-13-2010

Shuyang: http://paper.people.com.cn/jnsb/html/2007-09/13/content_19716392.htm, 09-13-2007

Lishui: http://www.jslszx.gov.cn/ReadNews.asp?NewsID=1712, 07-23-2010 many local governments have used the RSC program as an important tool to induce rural-urban migration.[32]

Specifically, in a media interview with the leader of the education department of Shou County in Anhui Province, the local official clearly indicated that he was instructed by the upper-level officials to "close schools in the old regions and move them to the new SEZ." The county government thus closed the best schools in the old region, although the RSC program was intended to close only those small and dilapidated schools.[33] There are plenty of similar records in other regions of China.

In a government report on the RSC program of Yulin City in Shaanxi Province, local officials found that among all the migrant workers, more than 30% chose to migrate because their children's needs for schooling could no longer be satisfied in their hometown, which suggests that access to education is an important factor that affects people's migration decisions in the context of rural China. This observation is also supported quantitatively by Xing (2014), who uses nationally representative datasets and finds that in villages where primary schools are closed, parents with school-aged children are much more likely to become migrant workers.

Identification Strategy

The identification strategy is based on a spatial-discontinuity design, as illustrated in Figure 2.1. Prefecture cities A and B are next to each other, there exists a SEZ on the border of the two prefecture cities, and it belongs to prefecture city A. Among the neighboring counties of this SEZ, some belong to the same prefecture city, A; others belong to the other prefecture city, B. We define those counties that are "in the neighborhood of the SEZ and belong to the same prefecture city" as "treated counties," and those that are "in the neighborhood of the SEZ but belong to a different prefecture city" as "control counties."

The treated and control counties are spatially adjacent to each other, so if the regional characteristics change smoothly over space, the treated and control counties should be balanced in these measures. Since the treated and control counties are all in the neighborhood of the same SEZ, they face the same spill-over effects from the development of this SEZ. Given that every prefecture city has a SEZ, it is also reasonable to assume that prefecture city A and prefecture city B are not systematically different.[34] Before the RSC program started, the provision of local public goods was determined at the county level, and neither the treated nor the control counties would internalize the positive externalities of their labor flowing to the SEZ, so the treated and the control counties would make the same competitive decisions: they would all provide high levels of public goods to prevent their labor from moving to the SEZ. All these reasons combined together, we expect that the treated and control counties should be well-balanced in every measure before the RSC program started.

---

[32] See examples in "Pay attention to the negative impacts of rural school consolidation," Guangming Daily, http://epaper.gmw.cn/gmrb/html/2015-05/24/nw.D110000gmrb_20150524_3-08.htm, 05-25-2015; "Misconceptions in the rural school consolidation policy," China Education Daily, http://www.jyb.cn/basc/xw/201109/t20110929_455976.html, 09-29-2011

[33] http://old.shouxian.gov.cn/contents/topic_view.php?id=169

[34] Nonetheless, we do provide direct tests in section 6.2 showing that even if there exists differences between cities A and B, such differences could hardly be driving the main results of the paper.

After the RSC program started, the leader of prefecture city A now makes a joint decision on primary schools for both the treated counties and the SEZ, so he internalizes the externalities associated with labor moving from the treated counties to the SEZ. In contrast, the leader of prefecture city B makes a decision on primary schools only for the control counties, so he does not internalize the externalities associated with labor moving from the control counties to the SEZ. That is to say, in terms of the provision of primary schools, there exists “coordination” between the treated counties and the SEZ, but “competition” between the control counties and the SEZ. Therefore, after the RSC program started, there should emerge a discontinuous gap between the treated counties and the SEZ, and the magnitude of this gap measures the causal effect of changing from a “competitive regime” to a “coordinative regime.”

Since the RSC program only applies to primary schools but not to any other types of public good, as a placebo, we should not observe the same magnitudes of impacts on other types of public good. Also, since politicians with different promotion incentives might have different objective functions, we expect to observe heterogeneity in competition and coordination among politicians with stronger promotion incentives.

2.3 Model

In this section, we formalize the aforementioned intuitions using a simple model of asymmetric fiscal competition, where labor is the mobile production factor, and governments are Leviathan (rent-maximizing). By comparing jurisdictions that make competitive decisions to those that make coordinated decisions, we derive our four main testable hypotheses.

Setup

This model modifies the classic models of income transfer competition (Wildasin, 1991) and asymmetric tax competition (Bucovetsky, 1991) to mimic our empirical setting illustrated in Figure 2.1. In our setup, as shown in Figure 2.2, we assume that there are three jurisdictions, indexed by 1, 2, and 3, corresponding to the “treated county,” the “SEZ,” and the “control county,” respectively.

The production function of jurisdiction *i* takes a quadratic form: $f_i(l_i) = \alpha_i \cdot l_i - \beta \cdot l_i^2$, where l_i stands for the population in jurisdiction *i*. By assumption, jurisdictions 1 and 3 are identical, while jurisdiction 2 is more productive: $\alpha_2 > \alpha_1 = \alpha_3$.[35]

Workers are freely mobile across jurisdictions. A worker in jurisdiction i earns a wage equal to the local marginal product of labor: $w_i = f_i'(l_i) = \alpha_i - 2 \cdot \beta \cdot l_i$. An individual in jurisdiction i consumes a numeraire good (equals to income w_i) and the local level of public good g_i, the utility function is linear and separable: $u_i = w_i + g_i$.

In equilibrium, every worker must be indifferent between the three jurisdictions:

$$f_1'(l_1) + g_1 = f_2'(l_2) + g_2 = f_3'(l_3) + g_3 \qquad (1)$$

and labor in all jurisdictions must add up to total population:

$$l_1 + l_2 + l_3 = L \qquad (2)$$

[35] This assumption means that if each jurisdiction has the same amount of labor, jurisdiction 2 would then have a higher marginal product of labor.

Equations 1 and 2 are similar to the conditions derived in the traditional tax competition models, and could fully characterize the allocation of labor among the three jurisdictions.

Combining equations 1 and 2, we derive the population of every jurisdiction as a function of both its own public good provision, and the levels of public good in the other two jurisdictions:

$$l_1 = \frac{L}{3} + \frac{(2 \cdot g_1 - g_2 - g_3) + (2 \cdot \alpha_1 - \alpha_2 - \alpha_3)}{6 \cdot \beta} \qquad (3)$$

$$l_2 = \frac{L}{3} + \frac{(2 \cdot g_2 - g_1 - g_3) + (2 \cdot \alpha_2 - \alpha_1 - \alpha_3)}{6 \cdot \beta} \qquad (4)$$

$$l_3 = \frac{L}{3} + \frac{(2 \cdot g_3 - g_1 - g_2) + (2 \cdot \alpha_3 - \alpha_1 - \alpha_2)}{6 \cdot \beta} \qquad (5)$$

The cost of providing public good g_i in jurisdiction i is $c(g_i) = c \cdot g_i^2$, and jurisdictional governments are Leviathan: they maximize their own rent, which equals to total production minus total wage and total cost of public good provision: $R_i = f_i(l_i) - f_i'(l_i) \cdot l_i - c(g_i)$. This rent could be understood as a general form of "government revenue."

Case 1: Each Jurisdiction Making its Own Decision (Pre-RSC)

Case 1 characterizes the situation before the RSC program starts: each jurisdiction chooses its own level of public good g_i to maximize its own rent, without internalizing any other jurisdiction's benefits. The three jurisdictions engage in fiscal competition, and each jurisdiction *i* solves:

$$\max_{g_i} R_i = f_i(l_i) - f_i'(l_i) \cdot l_i - c(g_i)$$

First order conditions give $l_i = 3 \cdot c \cdot g_i$ for each jurisdiction. Plugging equations (3), (4), and (5) into the three first order conditions, it is straightforward to show that:

$$g_2 > g_1 = g_3 \text{ and } l_2 > l_1 = l_3.$$

Case 2: Jurisdictions 1 and 2 Making a Joint Decision (Post-RSC)

Case 2 characterizes the situation after the RSC program starts: a planner is introduced to make a joint decision for jurisdictions 1 and 2 (corresponding to the leader of prefecture city A in the empirical setting), so he internalizes all the externalities caused by labor relocating between these two jurisdictions. Meanwhile, jurisdiction 3 still makes its own decision (corresponding to the leader of prefecture city B in the empirical setting), and still does not internalize the externalities of the outflow of its labor.

The planner of jurisdictions 1 and 2 chooses the levels of public good in these two constituencies to maximize the joint rent:

$$\max_{g_1, g_2} R_1 + R_2 = f_1(l_1) + f_2(l_2) - f_1'(l_1) \cdot l_1 - f_2'(l_2) \cdot l_2 - c(g_1) - c(g_2)$$

leading to the first order conditions:

$2 \cdot l_1 - l_2 = 6 \cdot c \cdot g_1$, and $2 \cdot l_2 - l_1 = 6 \cdot c \cdot g_2$

Jurisdiction 3 chooses the local level of public good to maximize its own rent:

$$\max_{g_3} R_3 = f_3(l_3) - f_3'(l_3) \cdot l_3 - c(g_3)$$

leading to the first order condition:

$$l_3 = 3 \cdot c \cdot g_3$$

Plugging equations (3), (4), and (5) into the three first order conditions, we obtain the reaction function (in terms of public good provision) of each jurisdiction. Combining the three reaction functions, we can show that:

$$g_2 > g_3 > g_1, \text{ and } l_2 > l_3 > l_1.$$

Extension: Other Public Goods

The RSC program only applies to primary schools, but not any other types of public good. To mimic this empirical feature accurately, in this section, we extend the baseline model to incorporate other types of public good, and explore how they would be affected by the RSC program.

In addition to g_i (primary schools), assume there are n other types of public good, and for computational tractability, assume that all the other public goods are identical, therefore the level of other public goods in jurisdiction i could be written as $n \cdot k_i$. Assume for simplicity that other types of public good are perfect substitutes of primary schools, the individual utility function then becomes $u_i = w_i + g_i + n \cdot k_i$, we could modify equations (1) and (2) to solve for the population of each jurisdiction at given levels of public goods.[36]

Given the extended setup, we could repeat the analysis of the cases in sections 3.2 and 3.3, to produce testable hypotheses for other public goods during the pre-RSC period and the post-RSC period, respectively.

Prior to the RSC program, each jurisdiction makes its own decision on g_i and k_i. Solving the maximization problem, given the symmetry assumptions for jurisdictions 1 and 3, $g_1 = k_1 = g_3 = k_3$ is trivially satisfied.

After the RSC program is implemented, a planner makes a joint decision on g_i for both jurisdictions 1 and 2, another planner makes a decision on g_3 only for jurisdiction 3; and each jurisdiction then separately makes its own decision on k_i. Solving the problem for the two planners and three jurisdictions altogether, we derive:

$$k_3 = \frac{l_3}{3 \cdot n \cdot c} > \frac{l_1}{3 \cdot n \cdot c} = k_1,$$

which leads to:

$$\frac{k_3 - k_1}{g_3 - g_1} = \frac{1}{(c-1) \cdot n}.$$

Same as in the baseline model, $g_3 - g_1 > 0$ still holds; but the sign of $k_3 - k_1$ is ambiguous: it is positive if $c > 1$, and negative if $c < 1$. However, so long as $c \neq 1$, when *n* becomes large enough, $\frac{k_3 - k_1}{g_3 - g_1}$ converges to zero.

Since in reality there are many different kinds of public good provided at the county level, it is likely that *n* is a large number, which means that for each specific type of public good other than primary schools, the impact of the RSC program is indirect and largely attenuated. Therefore, the difference in other public

[36] Note that the "identical" and "perfect substitution" assumptions are stronger than what we need, they are made to simplify computations. To derive the same qualitative implications, we only need to assume that "no other type of public good has specific substitutability or complementarity with primary schools."

goods between jurisdiction 1 (treated counties) and jurisdiction 3 (control counties) would likely remain minimal, even after the RSC program is implemented.

Extension: Political Incentives

In the baseline model, we assumed that all governments are rent-maximizing, which means there is no heterogeneity in political incentives. However, it is possible that when the leader of prefecture city A (B) has strong incentives for promotion, he will exert more efforts in coordination (competition), which means closing (opening) more primary schools in the treated (control) counties. To allow for such heterogeneity, in this section, we incorporate political incentives into our baseline model.

Specifically, we modify the objective function of a politician as a weighted average of rent and GDP:

$$\pi_i = (1-\theta)\cdot R_i + \theta \cdot Y_i = f_i(l_i) - (1-\theta)\cdot f_i'(l_i)\cdot l_i - (1-\theta)\cdot C(g_i)$$

where $\theta \in (0,1)$ is the measure of political incentives. When θ is larger, the politician cares more about promotion, so he puts a higher weight on GDP and a lower weight on rent, when making his decisions.

With this modified objective function, we repeat the optimizations of case 2 (section 3.3), under certain regularity conditions, we derive the comparative statics for public good provision:[37]

$$\frac{dg_1}{d\theta} < 0,\ \frac{dg_3}{d\theta} > 0.$$

Intuitively, these results indicate that, if the planner for jurisdiction 1 and jurisdiction 2 has strong promotion incentives, he would provide a lower level of public good in jurisdiction 1; if the planner for jurisdiction 3 has strong promotion incentives, he would provide a higher level of public good in jurisdiction 3.

Summary of Testable Hypotheses

In the empirical setting illustrated in Figure 2.1, the theoretical results could be summarized as four main testable hypotheses:

(1). Before the RSC program started, the treated counties and the control counties should be well balanced in every measure.

(2). After the RSC program started, the treated counties close more primary schools and lose more population than the control counties.

(3). After the RSC program started, except for primary schools, any other types of public good should remain balanced between the treated counties and the control counties.

(4). If the leader of prefecture city A has strong promotion incentives, there will be a higher level of school closure in the treated counties; if the leader of prefecture city B has strong promotion incentives, there will be a lower level of school closure in the control counties.

These four testable hypotheses are used to guide the empirical investigations.

[37] The cost of public good provision (c) must be larger than a lower bound. Due to length limits, details of this proof are not included in this draft, but are available upon request.

2.4 Data

We combine datasets from three different sources.

First, we have a novel dataset based on the "Township Conditions Survey (TCS)" conducted by the National Bureau of Statistics.[38] The TCS is a longitudinal survey conducted yearly since 2001, which covers all the 40906 townships in China. The TCS data includes a rich set of variables that are useful for our study, such as the number of schools, number of students, number of teachers, education expenditure, population, GDP, government expenditure, number of government officials, agricultural production, industrial output, etc. The TCS also provides detailed information on many different types of local public good, including hospitals, kindergartens, electricity, roads, water and sanitation, libraries, etc. The TCS is unique because it is the only dataset that has a complete coverage nationally at such a low administrative level, and to our best knowledge, this dataset has never been accessed for research before. By agreement, we have obtained access to the TCS data for all the townships in 20 major provinces from 2001 to 2011,[39] which perfectly overlaps with the timing of the Rural School Consolidation program. One potential caveat of the TCS data is that in some years it surveyed only some of the townships, while in other years it surveyed all of them, so selective attrition might be an issue. To address this concern, in addition to using the whole sample for analysis, in every table showing regression results, we present one column using only a balanced panel, and as we will see, the key results are essentially the same.

Second, we make use of rich information from the GIS maps of China. We first collect from administrative files a complete list of all the 1380 Special Economic Zones (SEZs) in China, as well as the type, area, level, and year of establishment for each of them. Then using GIS tools, we are able to pin down the geographic location of every SEZ, and the location of every county in China, and thus calculate the distance between every county-SEZ pair. Based on the GIS information, we could identify all the counties in the neighborhood of every SEZ, and categorize them into two types: those in the same city as the SEZ, and those in a different city as the SEZ. Matching this geographical information with the aforementioned TCS dataset constitutes the main dataset for this paper. The distribution of treated and control counties are shown in Figure 2.3, which are all the cases that follow the spatial discontinuity pattern illustrated in Figure 2.1.[40]

Third, to examine the fourth testable hypothesis of the model, we also make use of a prefecture city-level panel dataset on the Prefectural Party Secretaries (PPSs) of China from 2000 to 2010.[41] This dataset covers a total of 989 PPSs from 333 prefecture cities in 27 provinces, documenting detailed personal information including name, age, gender, nationality, education, experience, etc. Matching this dataset with the main dataset described above, we obtain an unbalanced panel covering 2737 townships from 2001 to 2010.

[38] The county level administrative dataset does not include information on primary schools, so our unit of analysis is at the township level, which is a smaller administrative unit than the county.

[39] There are 31 provinces, 2856 counties, 40906 townships in China.

[40] Some counties are in the neighborhood of a SEZ, but excluded from the sample, because they are urban districts, and are not affected by the RSC program.

[41] PPS is the highest level of government official at the city-level. For more details on the construction of the PPS dataset, see Chen (2015).

2.5 Empirics: Testing the Model

In this section, we formally examine the four testable hypotheses of our model: (1) before the RSC started, the treated and control townships should be balanced in every measure; (2) after the RSC started, the treated townships close more primary schools and lose more population than the control townships; (3) after the RSC started, except for primary schools, other types of local public good remain well-balanced between the treated and control townships; (4) when the leader of prefecture city A has strong promotion incentives, the treated townships close more primary schools, when the leader of prefecture city B has strong promotion incentives, the control townships retain more primary schools.

First Hypothesis: Balance Prior to RSC

Table 2.1 presents pre-treatment (before the RSC started) differences in means among key variables between the treated townships and the control townships. Following Imbens (2015), we use the "normalized differences" instead of the t-statistic as the measure of overlap,[42] which is independent of sample size.

The normalized difference is below 0.18 for all variables, below 0.12 for most of the variables, and even lower for the key outcome variables of interest. According to Imbens (2015), a normalized difference below 0.25 could be considered as "well balanced" between two groups. Hence, there is no statistically significant difference in any measure between the treated and control townships before the RSC started, and the key outcome variables (schools, students, teachers, population, employment) are especially well-balanced, supporting the first testable hypothesis of our model.

Second Hypothesis: Coordination Leads to More School Closure and Population Outflow than Competition

The second hypothesis indicates that after the RSC started, the treated townships should close more primary schools and lose more population. Specifically, since we expect that the people sorting after primary schools are mainly primary students, their parents, and primary school teachers, our empirical investigations focus on identifying the post-RSC differential trends between the treated and control townships in 5 key outcome variables: primary schools, primary students, primary school teachers, population, and employed population.

Econometric Model

The sample for analysis includes only townships in the neighboring counties of the borderline SEZs. Define the key variable $treatment_i$ as equal to 1 if township i lies in a treated county (in the same prefecture city as the neighboring SEZ), and 0 if township i lies in a control county (in a different prefecture city as the neighboring SEZ). The estimation equation is:

$$y_{ist} = \alpha \cdot treatment_i \cdot t + X'_{it} \cdot \beta + \lambda_{st} + \mu_i + \varepsilon_{ist} \qquad (6)$$

where y_{ist} is the outcome of interest in township i, in the neighborhood of SEZ s, in year t. X'_{it} is a vector of control variables measuring local characteristics, λ_{st} is SEZ-Year fixed effect, which restricts the variation being used for estimation to be within each SEZ neighborhood to increase the precision. μ_i is township fixed effect, ε_{ist} is the error term.

---

[42] Defined as difference in mean values over standard deviation: $(\mu_{xt} - \mu_{xc})/\sqrt{(s_{xc}^2 + s_{xt}^2)/2}$, where μ_{xt} and μ_{xc} are the sample averages of x in the treated and control groups, respectively; s_{xt}^2 and s_{xc}^2 are the sample variances of x in the treated and control groups, respectively.

Since the township data ranges from 2001 to 2011, which perfectly overlaps with the implementation of the RSC program, the coefficient α of the interaction term $treatment_i \cdot t$ would then capture the break in trends between the treated and control townships during the post-RSC period. According to our second testable hypothesis, α should be negative and significant, for all five key outcome variables (primary schools, primary students, primary school teachers, population, and employment).

To address concerns about spatial correlation among adjacent townships, we present standard errors clustered at the SEZ level; moreover, to address the concerns that non-adjacent townships within the same city could also be spatially correlated, we also present alternative standard errors that are two-way clustered at the City-SEZ levels (Cameron and Miller, 2015).[43]

Results

Table 2.2 presents the main results for primary schools. It shows that α is always negative and significant, which is consistent with our hypothesis: after the RSC started, because the treated townships coordinate with the SEZ, while the control townships compete with the SEZ, there emerges a negative gap in primary schools between the treated and control townships.

Column 1 controls for township fixed effect and year fixed effect, and indicates that on average a treated township closes 0.225 more primary schools every year than a control township. This could explain the closure of more than 13% of all primary schools during the 11 years that the RSC program was implemented.

Column 2 is the preferred specification, it further controls for SEZ-by-Year fixed effect. By doing so, we restrict the variation used for estimation to be within each SEZ cluster, thus could avoid making comparisons between treated and control observations that are actually far away from each other, which is a typical problem due to dimension reduction in many traditional Spatial-RD designs (Magruder, 2012). The magnitude of the estimated effect is even slightly larger than that of column 1: an average treated township closes an extra 2.9 primary schools in 11 years, which is more than 15% of the total amount of primary schools in an average township.

In column 3, we further include Province-by-Year fixed effect, and the point estimate is qualitatively the same, but smaller in magnitude. A potential explanation for the drop in magnitude is that there exists some level of cooperation between prefecture cities in the same province, and relatively less cooperation (or more fierce competition) between prefecture cities from different provinces, so the negative gap in public good provision should be even larger when prefecture city borders coincide with province borders.

Since our sample is an unbalanced panel, we are concerned that potential non-random attrition of townships in certain years might drive our results. In column 4, we keep only a balanced panel and run the same regression: the estimates are, if anything, even larger than that of the unbalanced panel. Therefore, the panel being unbalanced did not drive our main findings.

In column 5, we keep only those national level SEZs (selected by the central government, much larger than other SEZs), since those SEZs need more migrant workers and are also more politically rewarding for the city leaders, we expect the city governments to devote more efforts into these national SEZs, which would translate into a larger α. As we can see, the magnitude of estimated coefficient α doubles in column 5, which is consistent with our expectations.

---

[43] Each SEZ cluster includes all the townships in the neighboring county of this SEZ, each City cluster includes all the townships in that prefecture city.

In addition to differential trends in the number of primary schools, our model also predicts differential trends in local population. More specifically, in this empirical context, we expect the differential trends in population to be driven by differential trends in primary students, primary school teachers, and parents of school-aged children. And since the SEZs want to attract relatively skilled labor,[44] it is reasonable to assume that the migrating parents were originally employed back home.

We estimate the model (equation 6) with the same set of specifications for four additional outcome variables: primary students, primary school teachers, population, and employed population. In every case, we expect α to be negative and significant.

Table 2.3 presents the estimated coefficient α for these four outcome variables (results for primary schools are also kept for comparison), where each row uses a different outcome variable, each column uses a different specification, so every cell is an estimated coefficient from a separate regression. The overall patterns are highly consistent: for each of the outcome variables, and across different specifications, the estimated coefficient α tends to be negative and significant.

According to our preferred specification which controls for town fixed effect and SEZ-by-Year fixed effect (column 2), on average, a treated township loses 0.264 extra primary schools every year, which leads to an additional reduction of 3.76 teachers and 78 students. On average, a treated township has 173 additional out-migrants every year, among which 88 are employed. Notably, these estimates indicate a large treatment effect: between 2001 and 2011, the differential migration between the treated and control townships is as large as 7% of the total population of an average township in the sample.

In other specifications presented in Table 2.3, with township fixed effect always included, we explore models controlling for Year fixed effect (column 1), controlling for both SEZ-by-Year fixed effect and Province-by-Year fixed effect (column 3), using a balanced panel (column 4), the main results are highly robust. In column 5, when we keep only national SEZs, similar to the results for primary schools, the estimated effects on the other four outcome variables also become much larger.

Alternative Explanation

One potential concern about our results is that even if the treated and control townships are spatially adjacent, since they lie in different prefecture cities, it is still possible that there might be some other factors that make it relatively easier for people from the treated townships to migrate to the SEZ, resulting in higher outflows of population in the treated townships, including higher outflows of students. And as a result of more students leaving, schools in the treated townships become unnecessary and are thus closed.

Basically, this is a reverse causality concern: instead of the treated townships closing more schools to induce more out-migration, differential school closure might actually be the result of differential migration. One quick response to this argument would be: as presented in Table 2.1, the treated and control townships were very well balanced in every measure before the RSC started, including in students, population, and employment. This indicates that the presumption of "migration from the treated townships being easier" is unlikely to hold in the first place.

Nonetheless, we attempt to quantitatively rule out this reverse causality concern. To do so, we point out two critical features of the results in Table 2.3, which are consistent with our theoretical model, but inconsistent with the reverse causality explanation.

[44] They usually impose strict restrictions that do not allow the children of unemployed migrants to enroll in local schools.

The first observation is that if we divide $\alpha_{students}$ by $\alpha_{schools}$, we could conduct a rough back of the envelope calculation for the average size of the additional schools being closed in the treated townships. Using our preferred specification (column 2), for example, an additional school being closed in the treated townships has on average 295 students, significantly larger than the average school size in our sample (253 students). This pattern of closing large schools is inconsistent with the intentions of the RSC program (which requires the local governments to close the smallest schools); and inconsistent with the reverse causality explanation (which indicates that schools are closed because they do not have enough students); but only consistent with the incentive to induce migration proposed by our model.

The second observation is that, if there exists some other factors that cause differential migration between the treated and control townships, which in turn cause the observed difference in primary students, it is then highly likely that in addition to primary students, such factors should also have differential impacts on other similar population groups between the treated and control townships, i.e., we should expect differential migration of kindergarten kids, middle school students, high school students, college students, retired population, disabled population, etc. If any of these variables is also affected by these omitted factors, we would then expect that: $\alpha_{students} + \alpha_{employment} < \alpha_{population}$. However, according to our preferred specification, $\alpha_{students} + \alpha_{employment}$ (-166) is very close to $\alpha_{population}$ (-174), and the two numbers are statistically indistinguishable. Therefore, the reverse causality explanation is highly unlikely to be substantial.

Third Hypothesis: Other Public Goods are Unaffected

In this section, we formally examine the third testable hypothesis using the same econometric model (equation 6), but with other types of public good as outcome variables. Our model indicates that other types of public good would likely remain balanced between the treated and control townships even after the RSC started.

This hypothesis could also be perceived as a highly demanding placebo test, because it requires that if our main results are driven by any omitted variable, then whatever it is, this confounding factor must correlate and only correlate with primary schools, but not with any other types of local public good, not even middle schools or kindergartens.

Table 4 presents the results of this test. Among the 12 other types of public good measured in the data,[45] while all the coefficients are precisely estimated (with small standard errors), only one variable (number of villages that installed telephones) exhibits a differential trend between the treated and control townships, which is barely statistically significant at the 10% level. Moreover, the sign for this variable is positive, which is inconsistent with findings in the main results. All the 11 other types of public good remain very well balanced between the treated and control townships after the RSC started.

Therefore, the results of this test are consistent with our expectation, and strongly support the third testable hypothesis of our model.

Fourth Hypothesis: Promotion Incentives of Prefecture City Leaders Intensify Competition and Coordination

[45] Middle schools, kindergartens, number of villages with electricity, number of villages with phone signals, number of villages connected with paved roads, number of villages with TV signals, number of villages with tap water, total length of paved road, hospitals, libraries, cinemas, and sports stadiums.

In this section, we examine the fourth testable hypothesis derived from the extension our model: when the leader of prefecture city A has strong promotion incentives, the treated townships close even more primary schools, when the leader of prefecture city B has strong promotion incentives, the control townships close even fewer primary schools.

As well documented in the literature on Chinese meritocracy, other things being equal, the chance of promotion for a local official decreases discontinuously at a certain age threshold, and as a result, local officials generally devote much more efforts into boosting economic growth when they are below that critical age threshold (Li and Zhou, 2005). For politicians at the prefecture city level, the critical age threshold documented in the existing empirical literature generally falls in the [55, 57] interval (Yao and Zhang, 2011; Xi et al., 2015).

Therefore, we combine our main dataset with a panel dataset on the chief leaders of prefecture cities,[46] and follow the literature by defining a prefecture city leader as "incentivized" if his age is below 56.[47] The main estimation equation is:

$$y_{isct} = \alpha \cdot treatment_i \cdot t + \beta \cdot control_i \cdot incentive_{ct} + \gamma \cdot treatment_i \cdot incentive_{ct} + \lambda_{st} + \mu_i + \varepsilon_{isct}$$

where $incentive_{it}$ is a dummy variable that equals one if the leader of prefecture city c in year t is younger than 56, and zero otherwise. $control_i$ is a dummy variable that equals to 1 if township i lies in a control county, and zero otherwise; $treatment_i$ is a dummy variable that equals to 1 if township i lies in a treated county, and zero otherwise. y_{isct}, as before, represents the outcome of interest.

Intuitively, we could interpret β as the "effect of prefecture city B having an incentivized leader," and γ as the "effect of prefecture city A having an incentivized leader." Since incentivized leaders would have stronger incentives to promote economic growth, they are more likely to fight for labor, which intensifies fiscal competition and coordination. Therefore, we expect to see β being positive and significant (promotion incentives intensify competition), γ being negative and significant (promotion incentives intensify coordination).

Since in this section we are no longer focused on estimating and interpreting the average differential trends between the treated and control townships (coefficient α), we could now control more flexibly (less parametrically) for the differential dynamics between the two groups, which leads to the following specification:

$$y_{isct} = \beta \cdot control_i \cdot incentive_{ct} + \gamma \cdot treatment_i \cdot incentive_{ct} + \omega_{\mathrm{Tt}} + \lambda_{st} + \mu_i + \varepsilon_{isct}$$

where ω_{Tt} is Treatment-by-Year FE, it fully absorbs the differential dynamics between the treated and control groups, and would therefore improve the precision of estimation for political incentives (coefficients β and γ).

As shown in Table 2.5, the results are exactly as we expected, and they have huge magnitudes: according to the preferred flexible specification (column 4), having an incentivized leader in prefecture city A means a treated township closes an extra 0.66 primary schools every year, while having an incentivized leader in prefecture city B means a control township retains an extra 0.97 primary schools every year. The results are qualitatively robust to using less flexible specifications (columns 1 and 2),

[46] Prefecture Party Secretaries. We drop the "outliers" that are younger than 40, which is smaller than 1% of the sample.

[47] Our results are qualitatively robust to slightly different choices of this threshold.

using only a balanced panel (column 3), and controlling for personal characteristics of prefecture city leaders (column 5).

Therefore, more incentivized politicians exert more efforts in fiscal competition and coordination, confirming the fourth prediction of our model. In addition, the opposite signs and comparable magnitudes of β and γ also suggest that both "coordination" and "competition" are important in producing the discontinuous gaps in the key outcome variables between the treated and control townships.

2.6 Additional Results

In this section, we present three additional results: section 6.1 investigates the impacts of fiscal competition and coordination on the quality of education; section 6.2 introduces an alternative empirical strategy to confirm that our results are not driven by underlying differences between adjacent prefecture cities; section 6.3 presents robustness checks using only SEZs that were established before the RSC program started.

Quality of Education

The discussions of this paper have mainly focused on the quantity of education (number of primary schools). However, for prefecture city A (B) to induce high-skilled labor to migrate (prevent high-skilled labor from migrating), the quality of education could also be important. As discussed in section 2.3, qualitative evidence confirms that some prefecture cities did close their best schools and relocate them to their SEZs, suggesting that the quality of education also plays an important role in local competition and coordination.

We proxy for the quality of education with "education expenditure per student," which is widely used in the economics of education literature (Jackson et al., 2015). However, unlike the key outcome variables used before (schools, students, teachers, population, employment), this variable is only available after 2006, because of a national reform on educational finance happened at that time, which required the local governments to keep detailed electronic records of educational expenditures.[48]

Adopting our main specification (equation 6) with "education expenditure per student" as the outcome variable, we could test whether the quality of education has differential trends between the treated and control townships. If it does, we would expect α to be negative and significant.

The results are presented in Table 2.6. As expected, from 2007 to 2011, the gap in average education expenditure between the treated and control townships increases at a rate of 156 Yuan per year, more than 8% of the education expenditure per student in the sample. However, unlike previous findings on other outcome variables, this result does not hold for national SEZs, which might be due to the fact that data for this variable is only available after 2006, by which time the quality adjustment in the national SEZs might has already been achieved.

Worth noting is that other than our proposed mechanism, where prefecture city A coordinates between the treated townships and the SEZ to induce migration (and prefecture city B competes against the SEZ to prevent labor from leaving the control townships), this finding on the quality of education could hardly be reconciled with other confounding explanations. For instance, if the aforementioned reverse causality hypothesis holds (migration happened before school closure), then there is a reason to close schools after people leave, but there is no reason to also lower the quality of education in those local schools that are

[48] The "New Mechanisms Reform," see Wang (2008) for more details.

not being closed. Also, were it not for the coordinative incentive, one might even expect that high fiscal income generated by the SEZ could create positive spill-over effects for the treated townships, and thus lead to an improvement of education quality in the treated townships, rather than the other way round.

Prefecture City Heterogeneity

As discussed in section 2.1, by 2006, every prefecture city in China has at least one SEZ. So in our empirical setting, although the variation for identification relies on the fact that prefecture city A has a SEZ lying on its border, prefecture city B also has its own SEZ, albeit not lying on the border shared with prefecture city A. Therefore, there is no reason to be concerned about the possibility that if prefecture city A has a SEZ but prefecture city B does not have one, they might be systematically different in many unobservable ways, which drive the results we found.

However, one might still be concerned about the fact that prefecture city A's SEZ lies on its border shared with prefecture city B, while prefecture city B's SEZ does not lie on the same border. If such difference in locational choices of SEZs could reflect some underlying differences between the two prefecture city governments, for instance, different philosophies in decision-making or different abilities in lobbying, it is then possible that such differences might be omitted from the main specification, and could potentially drive the results of differential school closure between the treated and control townships.

To address this concern, we adopt an alternative empirical strategy that avoids the potential selection issues at the prefecture city level. The setting for this strategy is illustrated in Figure 2.4. As shown, we identify all the cases where prefecture city C has a SEZ on its border that is not shared with prefecture city D, and prefecture city D has a SEZ on its border that is shared with prefecture city C. We define the counties that are in prefecture city C and in the neighborhood of prefecture city C's SEZ as "treated," and those counties that are in prefecture city C but in the neighborhood of prefecture city D's SEZ as "control." Since the treated and control groups are all in the neighborhoods of SEZs, they are expected to have comparable "natural" trends for urbanization. More importantly for this exercise, since both the treated and control townships are in prefecture city C, the decision of school closure for both groups are all decided by the same agent: the leader of prefecture city C. Therefore, if the results in section 5.2 are mainly driven by selection at the prefecture city level, or any other omitted confounding differences between the two adjacent prefecture cities, we would expect that the treated and control townships in this alternative setting no longer have any difference in school closure. However, if our proposed mechanism is correct, prefecture city C would want to close more schools in the treated townships to induce migration to its own SEZ, but retain more schools in the control townships to keep its labor. In that case, we would expect to find the same pattern as before: the number of schools decreases faster in the treated townships than in the control townships.

For estimation, we simply adopt the specification of equation 6, but replace the SEZ-by-Year fixed effect with City-by-Year fixed effect. Then, if selection at the prefecture city level is really an important concern, the coefficient α should be statistically indistinguishable from zero; if our proposed mechanism is correct, α should be negative and significant.

Table 2.7 presents the results. In column 1, we control for township fixed effect and year fixed effect, and the estimated coefficient is -0.198, similar to what we obtained using the main identification strategy in the whole sample (-0.225). Column 2 shows the preferred specification where we further control for City-by-Year fixed effect. The coefficient is negative and significant, and almost doubled in magnitude. This strongly rejects the "city heterogeneity" hypothesis, but could be explained by competition and coordination between local governments. Column 3 uses only a balanced panel, and the results go through, with even larger magnitudes.

Compared to our main identification strategy, a potential caveat for this alternative strategy is that we no longer have the unique spatial-discontinuity feature: the treated and control townships are now in the neighborhoods of different SEZs. Therefore, one might worry about the size of SEZs being a confounding factor: if for some reason the SEZ in prefecture city C is larger than the SEZ in prefecture city D, then even absent of differential incentives (to coordinate or to compete), prefecture city C would still naturally close more primary schools in the treated townships than the control townships, which leads to an over-estimation of the coefficient α. In column 4, we attempt to address this concern by controlling for the interaction term of the "area of a SEZ" and a continuous time variable, which absorbs the differential trends caused by SEZ size. As we can see, the interaction term has a negative coefficient, meaning that more schools are closed in the neighborhood of a larger SEZ. After we control for such impacts of SEZ size, the estimated gap between the treated and control townships shrinks to -0.219, which is statistically significant at the 5% level, and is extremely close to our baseline result using the main identification strategy (-0.225).

As suggested by this alternative empirical strategy, it is unlikely that our results are driven by underlying differences between adjacent prefecture cities, instead, it is highly likely that the proposed mechanism (coordination against competition) is dominant in explaining our main findings.

SEZs Established after the RSC Started

Another potential concern comes from the fact that some SEZs were established after the start of the RSC program. Specifically, the RSC program provided a prefecture city government with the ability to coordinate between its own counties and its own SEZ, and we might expect that the prefecture city government would take this factor into consideration when they make locational choices for new SEZs after the RSC started. While this possibility does not necessarily contradict the hypotheses of competition and coordination, it might be a concern if our previous regression results were partly capturing the strategic locational choices of SEZs that were established after 2002. To address this issue, in this section, we replicate our main results (Table 2.3) using only the subsample of SEZs that were established before 2002.[49]

In Table 2.8, we present the estimated coefficients for the five key outcome variables: primary schools, primary students, primary school teachers, population, and employed population. The results are overall very similar to that presented in Table 2.3, and if anything, slightly larger in magnitude. The estimated standard errors are also slightly larger, potentially due to the reduction in sample size by excluding those SEZs established after 2002. This suggests that our main results are not driven by the strategic locational choices of SEZs after the RSC started.

2.7 Conclusion

In this paper, we identify causal evidence on fiscal competition and coordination among local governments.

[49] Among the 58 SEZs satisfying our empirical setting of Figure 2.1, 11 were established after 2002, and those are excluded from the subsample used in this section.

We first introduce a unique empirical context, which is created by the combination of two national programs in China: the "Special Economic Zones (SEZ)" program and the "Rural School Consolidation (RSC)" program. In a spatial discontinuity setting, we identify all the SEZs on the prefecture city borders, and for their neighboring counties, we define those that lie in the same prefecture city as "treated," and those that lie in the different prefecture city as "control."

We present a simple model of asymmetric fiscal competition which mimics the empirical setting, and formally derive four testable hypotheses: (1) before the RSC program started, the treated counties and the control counties should be well balanced in every measure; (2) after the RSC program started, the treated counties close more primary schools and lose more population, as compared to the control counties; (3) after the RSC program started, except for primary schools, any other types of public good would likely remain balanced between the treated counties and the control counties; (4) if the leader of prefecture city A has strong promotion incentives, the treated counties would close even more primary schools, if the leader of prefecture city B has strong promotion incentives, the control counties would close even fewer primary schools.

Applying the spatial discontinuity design to a unique panel dataset of Chinese townships, we find empirical evidence strongly supporting all four testable hypotheses. Further analysis provides three additional results: (1) the quality of education also follows the same pattern of competition and coordination; (2) the main results are not driven by underlying differences between adjacent prefecture cities; (3) the main results are not driven by strategic establishment of SEZs after the RSC program started. Overall, the empirical findings provide causal evidence on fiscal competition and coordination among local governments in China, which could also add to the more general literature on fiscal federalism and decentralization.

Finally, our results also have important policy implications: when designing centralized programs, to avoid unexpected consequences, the strategic interactions among local governments should be systematically taken into consideration.

Figure 2.1 Empirical Setting

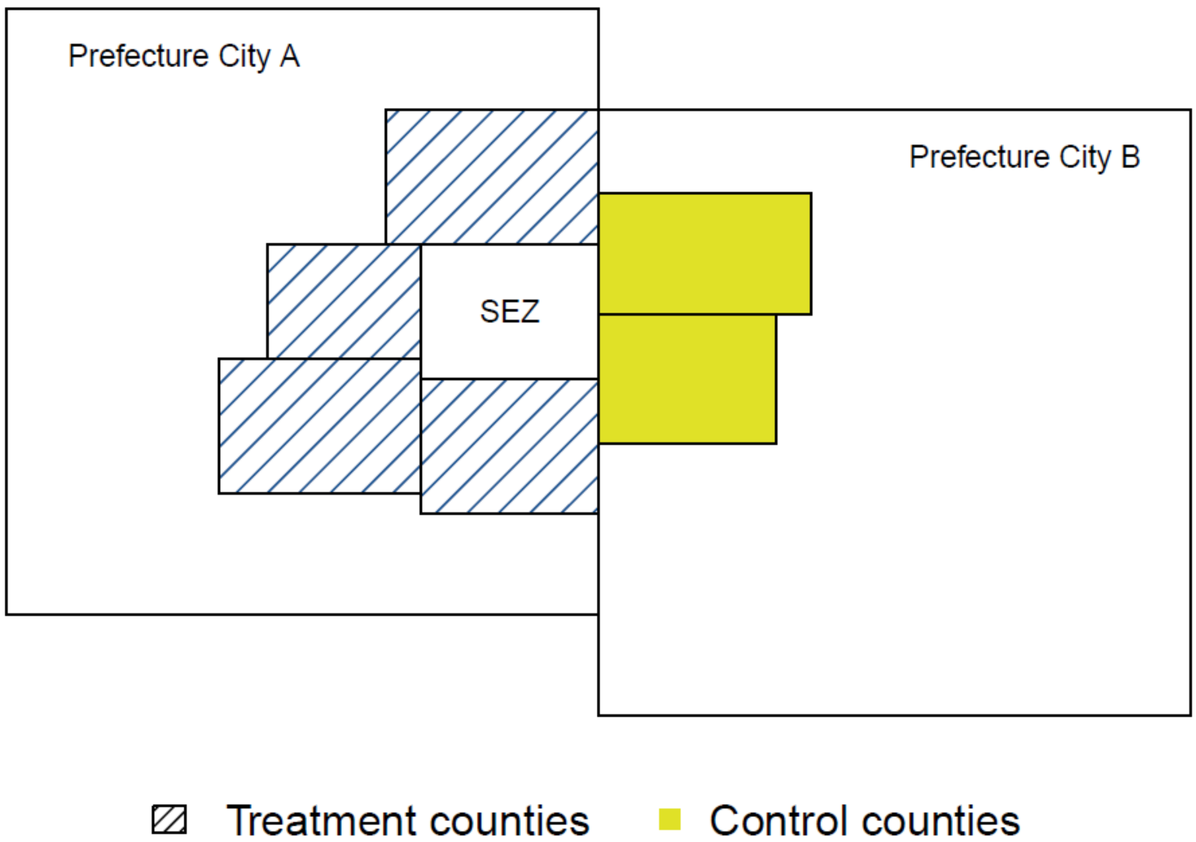


Figure 2.2 Theoretical Setting

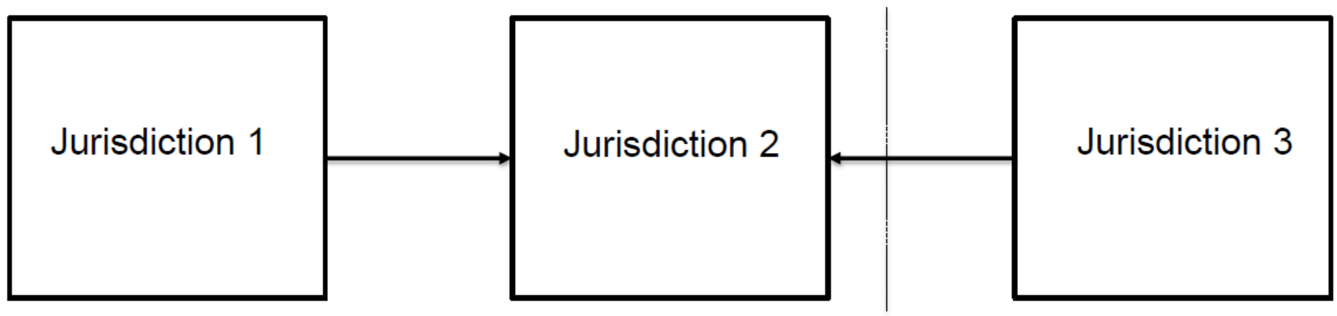


Figure 2.3 Sample Distribution

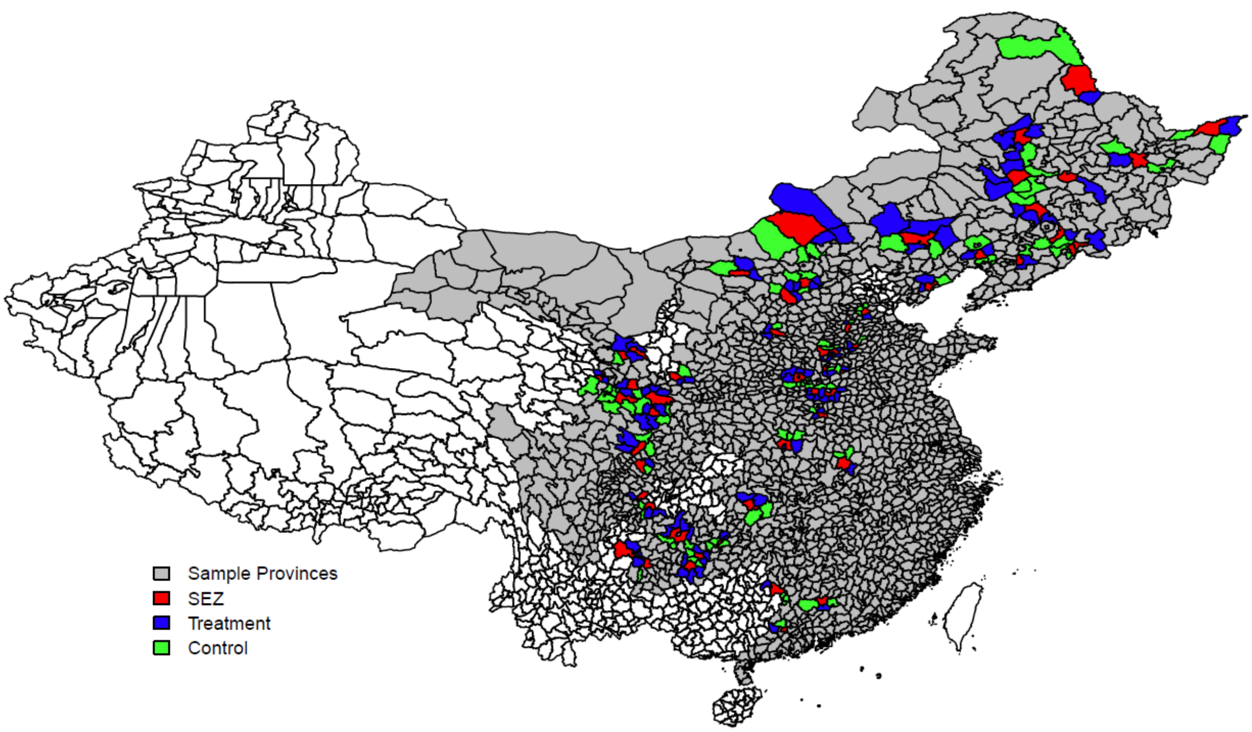


Figure 2.4 Alternative Empirical Setting

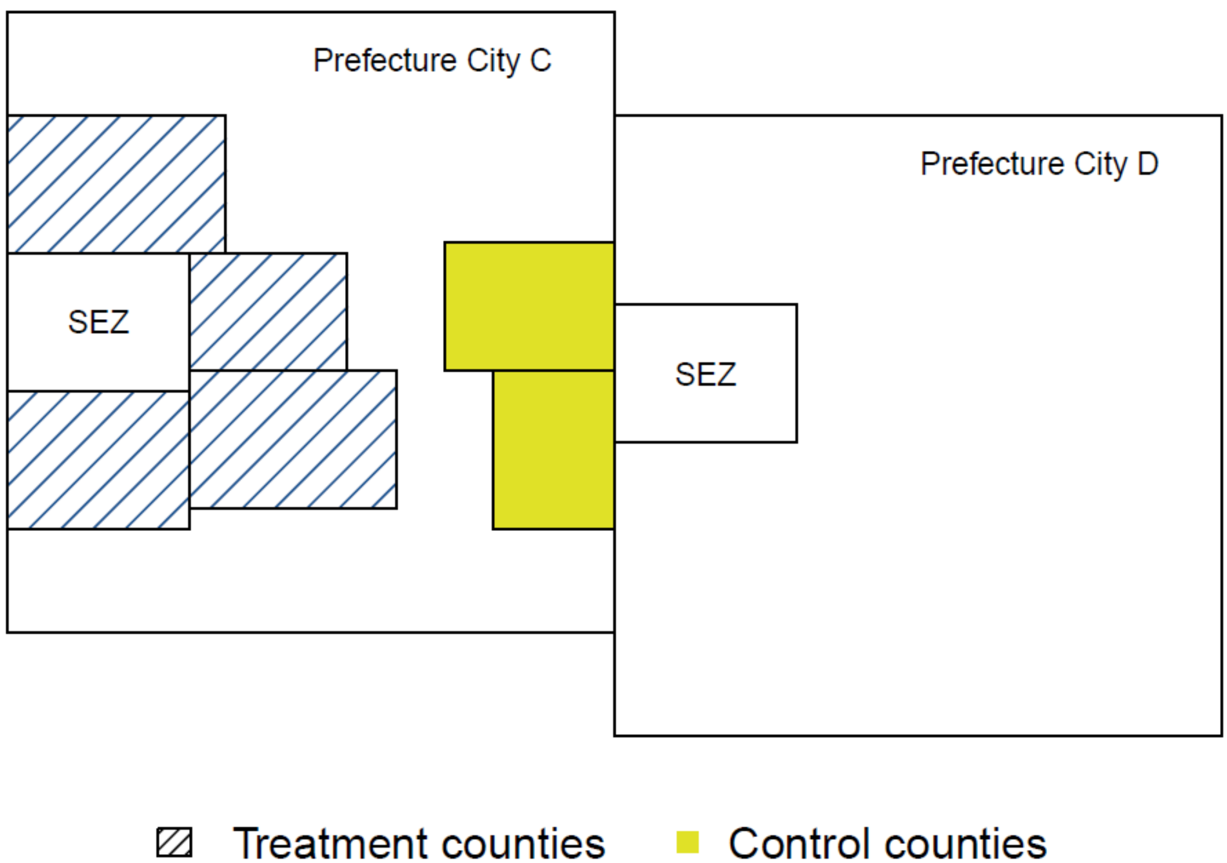


Table 2.1 Pre-RSC Balance Test

	Control		Treated		Normalized Difference
VARIABLES	**Mean**	**S.D.**	**Mean**	**S.D.**	
Primary Schools	18.85	8.84	19.35	11.87	0.05
Primary Students	4619.27	3467.60	4904.75	3891.52	0.08
Primary Teachers	272.04	210.39	297.39	240.69	0.11
Population	28346.54	18125.95	30350.22	19263.48	0.11
Employed Population	14146.94	10305.99	15169.82	11000.17	0.10
Income	2041.27	772.75	1997.45	849.69	0.05
Area	16263.93	32480.79	17307.90	28386.50	0.03
Number of Communities	1.95	5.23	2.21	7.23	0.04
Number of Villages	16.93	9.79	18.45	10.13	0.15
Electrified Villages	16.81	10.18	18.51	11.40	0.16
Villages w\ Telephones	15.17	10.18	15.92	11.59	0.07
Villages w\ Roads	15.52	9.88	17.42	11.30	0.18
Villages w\ TV	7.65	8.79	8.26	9.78	0.07
Villages w\ Tap Water	8.22	9.89	8.90	9.71	0.07
Villages w\ Incinerators	0.26	0.72	0.30	1.78	0.03
Number of Hospitals	5.38	10.46	5.49	12.35	0.01
Number of Kindergartens	6.22	20.14	6.87	9.93	0.04
Number of Libraries	1.31	2.06	1.85	5.59	0.13
Number of Stadiums	0.37	1.94	0.28	1.11	0.06
Observations	662		729		

Notes: This table reports the summary statistics by treatment and control townships in 2001 (pre-RSC). Following Imbens (2015), we use Normalized Difference to measure the overlap between the two groups, and all the variables are found to be well balanced. Two measures of public good (Number of Middle Schools, Number of Cinemas) are not included in this table because there are not available in the 2001 TCS dataset.

Table 2.2 Main Strategy: Primary Schools

Dependent Variable:	Number of Primary Schools				
	(1)	(2)	(3)	(4)	(5)
Treatment*Year	-0.225***	-0.264***	-0.155***	-0.223***	-0.470***
	(0.0429)	(0.0414)	(0.0416)	(0.0692)	(0.108)
	[0.142]	[0.0918]	[0.0620]	[0.0589]	[0.0986]
Township FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	No	No	No	No
SEZ-Year FE	No	Yes	Yes	Yes	Yes
Province-Year FE	No	No	Yes	Yes	Yes
R-squared	0.343	0.477	0.513	0.536	0.457
Observations	20,774	20,774	20,774	8,206	2,390

Notes: This table reports the post-RSC differential trends in primary schools between the treated and control townships. Column 1 controls for township FE and year FE, column 2 (preferred specification) controls for township FE and SEZ-by-Year FE, column 3 further controls for Province-by-Year FE, column 4 uses only a balanced panel, column 5 uses only the townships that are adjacent to national-level SEZs. Standard errors are clustered at the SEZ level; ***, **, and * indicate significance at 1%, 5% and 10% levels. Alternative standard errors (two-way clustered at the city and SEZ levels) are also presented.

Table 2.3 Main Strategy: Key Outcomes

	(1)	(2)	(3)	(4)	(5)
$\alpha_{schools}$	-0.225***	-0.264***	-0.155***	-0.223***	-0.470***
	(0.0429)	(0.0414)	(0.0416)	(0.0692)	(0.108)
	[0.142]	[0.0918]	[0.0620]	[0.0589]	[0.0986]
$\alpha_{students}$	-45.85**	-77.76***	-67.42***	-56.96	-250.4***
	(19.53)	(23.49)	(25.33)	(48.90)	(77.06)
	[40.97]	[27.85]	[31.53]	[44.39]	[96.35]
$\alpha_{teachers}$	-1.779	-3.756**	-3.291**	-3.381	-7.728***
	(1.182)	(1.578)	(1.530)	(2.891)	(2.964)
	[2.214]	[1.631]	[1.763]	[2.281]	[2.843]
$\alpha_{population}$	-115.0*	-173.5**	-210.0***	-241.5*	-909.4***
	(63.27)	(76.69)	(80.40)	(145.5)	(286.8)
	[105.5]	[103.5]	[75.19]	[137.0]	[233.1]
$\alpha_{employment}$	11.10	-87.81*	-124.1**	-156.4*	-566.2***
	(41.78)	(48.40)	(52.68)	(93.36)	(167.9)
	[62.96]	[68.70]	[58.57]	[91.11]	[42.27]

Notes: This table reports the post-RSC differential trends in all the five key outcome variables between the treated and control townships. Each cell presents an estimated coefficient obtained from a separate regression, with each row using a different outcome variable, and each column using a different specification: column 1 controls for township FE and year FE, column 2 (preferred specification) controls for township FE and SEZ-by-Year FE, column 3 further controls for Province-by-Year FE, column 4 uses only a balanced panel, column 5 uses only the townships that are adjacent to national-level SEZs. Standard errors are clustered at the SEZ level; ***, **, and * indicate significance at 1%, 5% and 10% levels. Alternative standard errors (two-way clustered at the city and SEZ levels) are also presented

Table 2.4 Other Public Goods

Dependent Variables:	**Middle Schools**	**Kindergarten**	**Electric Villages**	**Phone Villages**	**Road Villages**	**TV Villages**	**Water Villages**	**Road Length**	**Hospitals**	**Libraries**	**Cinema**	**Stadiums**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment*Year	0.00285	-0.0510	0.0372	0.113*	-0.00751	0.0318	0.0124	0.451	0.00008	-0.0175	0.00125	-0.0041
	(0.00898)	(0.0518)	(0.0327)	(0.0677)	(0.0392)	(0.0557)	(0.0398)	(0.655)	(0.0004)	(0.0233)	(0.00260)	(0.0179)
	[0.0105]	[0.0788]	[0.0668]	[0.0823]	[0.0762]	[0.102]	[0.0474]	[1.101]	[0.0004]	[0.0236]	[0.00199]	[0.0105]
Township FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SEZ-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.421	0.198	0.762	0.743	0.726	0.602	0.594	0.330	0.070	0.209	0.078	0.142
Observations	16,656	20,770	20,774	20,774	20,774	20,773	20,773	20,774	20,774	20,765	16,570	20,707

Notes: This table reports the parallel post-RSC trends in 12 different measures of local public good using the preferred specification. Township FE and SEZ-by-Year FE are included in every specification. Standard errors are clustered at the SEZ level; ***, **, and * indicate significance at 1%, 5% and 10% levels. Alternative standard errors (two-way clustered at the city and SEZ levels) are also presented.

Table 2.5 Political Incentives

Dependent Variable:	Number of Primary Schools				
	(1)	(2)	(3)	(4)	(5)
Incentive*Control	0.423	1.012**	1.403**	0.974**	1.780***
	(0.367)	(0.405)	(0.573)	(0.410)	(0.529)
	[0.418]	[0.450]	[0.503]	[0.461]	[0.444]
Incentive*Treatment	-0.349	-0.580*	-0.478	-0.663**	-0.794**
	(0.293)	(0.304)	(0.432)	(0.307)	(0.331)
	[0.335]	[0.361]	[0.390]	[0.360]	[0.374]
Treatment*Year	-0.360***	-0.0677	-0.178**		
	(0.0671)	(0.0594)	(0.0869)		
	[0.175]	[0.0843]	[0.0738]		
Treatment-by-Year FE	No	No	No	Yes	Yes
PPS Characteristics	No	No	No	No	Yes
Township FE	Yes	Yes	Yes	Yes	Yes
SEZ-by-Year FE	Yes	Yes	Yes	Yes	Yes
Province-by-Year FE	No	Yes	Yes	Yes	Yes
R-squared	0.439	0.486	0.487	0.486	0.498
Observations	13,074	13,074	5,981	13,074	11,338

Notes: This table reports the effects of the political incentives of prefecture city leaders on the post-RSC differential trends in primary schools between the treated and control townships. Column 1 controls for Township FE and SEZ-by-Year FE, column 2 further controls for Province-by-Year FE, column 3 uses only a balanced panel, column 4 replaces the interactions of treatment and year with Treatment-by-Year FE which is more flexible, column 5 further controls for personal characteristics of the Prefectural Party Secretary (gender, nationality, years of working experience, political faction, previous position). Standard errors are clustered at the SEZ level; ***, **, and * indicate significance at 1%, 5% and 10% levels. Alternative standard errors (two-way clustered at the city and SEZ levels) are also presented.

Table 2.6 Quality of Education

Dependent Variable:	Education Expenditure per Student				
	(1)	(2)	(3)	(4)	(5)
Treatment*Year	0.00147	-0.0156***	-0.0149***	-0.0186***	0.0146
	(0.00624)	(0.00415)	(0.00447)	(0.00640)	(0.0135)
	[0.0132]	[0.00733]	[0.00644]	[0.00915]	[0.0170]
Township FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	No	No	No	No
SEZ-Year FE	No	Yes	Yes	Yes	Yes
Province-Year FE	No	No	Yes	Yes	Yes
R-squared	0.054	0.378	0.391	0.197	0.232
Observations	4,973	4,973	4,973	2,266	489

Notes: This table reports the post-2006 differential trends in education expenditure per student. Column 1 controls for township FE and year FE, column 2 (preferred specification) controls for township FE and SEZ-by-Year FE, column 3 further controls for Province-by-Year FE, column 4 uses only a balanced panel, column 5 uses only the townships that are adjacent to national-level SEZs. Standard errors are clustered at the SEZ level; ***, **, and * indicate significance at 1%, 5% and 10% levels. Alternative standard errors (two-way clustered at the city and SEZ levels) are also presented.

Table 2.7 Alternative Strategy

Dependent Variable:	**Number of Primary Schools**			
	(1)	(2)	(3)	(4)
Treatment*Year	-0.198*	-0.386***	-0.665***	-0.219**
	(0.116)	(0.104)	(0.234)	(0.0902)
	[0.350]	[0.207]	[0.295]	[0.134]
SEZ Area*Year				-0.0011***
				(0.00021)
				[0.00036]
Township FE	Yes	Yes	Yes	Yes
Year FE	Yes	No	No	Yes
City-by-Year FE	No	Yes	Yes	No
R-squared	0.280	0.440	0.505	0.454
Observations	3,674	3,674	847	3,674

Notes: This table reports the post-RSC differential trends in primary schools using the alternative specification. Columns 1 controls for township FE and Year FE, columns 2, 3, and 4 control for township FE and City-by-Year FE. Column 3 uses only a balanced panel. Column 4 controls for the linear trend caused by the size (area measured in square kilometers) of the neighboring SEZ. Standard errors are clustered at the SEZ level; ***, **, and * indicate significance at 1%, 5% and 10% levels. Alternative standard errors (two-way clustered at the city and SEZ levels) are also presented.

Table 2.8 Robustness Check

	(1)	(2)	(3)	(4)	(5)
$\alpha_{schools}$	-0.271***	-0.292***	-0.175***	-0.249***	-0.470***
	(0.0463)	(0.0445)	(0.0447)	(0.0736)	(0.108)
	[0.155]	[0.0980]	[0.0632]	[0.0620]	[0.0986]
$\alpha_{students}$	-52.74**	-102.4***	-96.39***	-95.08*	-250.4***
	(22.50)	(26.40)	(28.79)	(55.40)	(77.06)
	[42.83]	[28.45]	[32.74]	[48.65]	[96.35]
$\alpha_{teachers}$	-1.898	-4.687***	-4.347**	-4.718	-7.728***
	(1.391)	(1.806)	(1.768)	(3.340)	(2.964)
	[2.236]	[1.793]	[1.949]	[2.490]	[2.843]
$\alpha_{population}$	-166.2**	-178.2**	-222.8**	-225.7	-909.4***
	(71.74)	(85.29)	(90.71)	(163.4)	(286.8)
	[113.1]	[116.4]	[85.71]	[156.6]	[233.1]
$\alpha_{employment}$	-22.08	-103.6*	-143.9**	-167.5	-566.2***
	(47.50)	(53.98)	(59.75)	(105.6)	(167.9)
	[69.70]	[76.81]	[66.39]	[103.4]	[42.27]

Notes: This table replicates table 3 using only the subsample of SEZs that were established before 2002. Each cell presents an estimated coefficient obtained from a separate regression, with each row using a different outcome variable, and each column using a different specification: column 1 controls for township FE and year FE, column 2 (preferred specification) controls for township FE and SEZ-by-Year FE, column 3 further controls for Province-by-Year FE, column 4 uses only a balanced panel, column 5 uses only the townships that are adjacent to national-level SEZs. Standard errors are clustered at the SEZ level; ***, **, and * indicate significance at 1%, 5% and 10% levels. Alternative standard errors (two-way c

3 Influence Activities and Bureaucratic Performance: Evidence from a Large-Scale Field Experiment in China

3.1 Introduction

For a large share of jobs in a modern economy, objective performance measures are difficult to obtain, leading employers to rely heavily on supervisor's subjective evaluations to provide work incentives (Prendargast, 1999; Deb et al., 2016). This is particularly ubiquitous in the public sector, due to the inherent problems of measurability and multiplicity of the goals for most civil service jobs (Olken and Pande, 2013; Finan et al., 2015).

While subjective performance measures could potentially improve contractual power (Gibbons and Murphy, 1992; Baker et al., 1994), they also open the door to influence activities: employees can take actions to affect the evaluator's decision in their favor, which might be against the interests of the organization (Milgrom and Roberts, 1988; Milgrom, 1988). For the organization, influence activities thus create a tradeoff between taking advantage of the supervisor's local information and letting the supervisor manipulate decisions in his own interests, undermining the organization's ability to fully incentivize the agents to perform on their jobs. However, despite being consistent with anecdotal examples and case studies, the long-standing theoretical discussion on the existence and implications of influence activities is yet to be backed up by rigorous empirical evidence (Oyer and Schaefer, 2011; Lazear and Oyer, 2012).

Empirically investigating influence activities is fundamentally challenging due to at least three reasons. First, behaviors such as buttering up supervisors or providing personal favors to them are difficult to observe by nature. In fact, agents may try their best to hide such behaviors because they are usually deemed as inappropriate. Second, even if these behaviors are observed, it is difficult to conclude that they are driven by intentions to improve evaluation outcomes (instead of simply being friendly), making us unable to exclusively classify them as influence activities. Third, even if we establish the existence of influence activities, without being able to exogenously vary such behaviors across agents, we are still unable to quantify the causal impacts of such behaviors on work performance.

In this paper, we conduct a large-scale field experiment, which aims to address these three challenges altogether, and provide the first rigorous empirical evidence on the existence and implications of influence activities in workplace. Our experiment focuses on China's "3+1 Supports" program, which hires more than 30,000 college graduates annually to work in rural township governments on two-year contracts. These junior state employees are referred to as College Graduates Civil Servants (CGCSs) in this paper.

An important institutional feature of the Chinese governance system is its dual-leadership arrangement (Shirk, 1993): within each government organization/subsidiary, there are always two leaders: a "party leader" (i.e. party secretaries at various levels) and an "administrative leader" (i.e. head in a village, mayor in a city).[50] As a result of this duality system, each CGCS is supervised by two senior leaders, both assign job tasks and

[50] The two leaders have large overlap in their responsibilities, introducing *de facto* checks and balances. See Li (2018) for more detailed discussions on the institutional details of the duality system.

provide feedbacks on a regular basis. In the *status quo,* each CGCS is evaluated by one of his two co-supervisors every year, and the evaluation outcome in turn determines whether the CGCS can be promoted to a "tenured" position upon completing his two-year term, which is a highly sought-after outcome for most CGCSs due to the prestige of civil servant jobs in China.

Exploiting this unique setting, we collaborate with two provincial governments in China to randomize two different types of performance evaluation schemes among their 3,785 CGCSs. In both schemes, we randomly select one of the two supervisors to be the "evaluator" (whose opinion matters for future promotion), but also collect the assessment from the other non-evaluating supervisor. The only difference is that, in the first "revealed" scheme, we announce the identity of the evaluator to the CGCS at the beginning of the evaluation cycle, so that throughout the year, the CGCS knows whose opinion is 100% responsible for his career development; while in the second "masked" scheme, we keep the identity of the evaluator secret until the end of the evaluation cycle, so that throughout the year, the CGCS perceives each supervisor to be having a 50% chance of determining his promotion.

We find that, in the revealed scheme, the randomly selected evaluating supervisor gives significantly more positive assessments of CGCS performance than his non-evaluating counterpart. In addition, when we ask the colleagues of the CGCS to speculate who among the two supervisors would give higher assessments of CGCS performance, they are also more likely to (correctly) guess the randomly selected evaluator, of which they were not informed. These results are consistent with a scenario where the agent is able to impose supervisor-specific influence activities to improve evaluation outcomes, and such influence activities could partly be observed by his other co-workers.

In the masked scheme, since the CGCS no longer knows who the evaluating supervisor is during the evaluation cycle, the asymmetry in supervisor assessments disappears. Furthermore, the masked scheme improves CGCS performance, as measured by both average colleague assessment and average supervisor assessment. The effect of the masked scheme is stronger when the CGCS is more risk averse, when the two supervisors have more aligned preferences for performance, and when the two supervisors have little information asymmetry. These results can be rationalized in a conceptual framework, where introducing uncertainty in the identity of the evaluator incentivizes the agent to reallocate efforts from evaluator-specific influence activities toward productive dimensions that would be appreciated by both supervisors.

We conduct a battery of additional analysis to rule out alternative interpretations of our baseline findings, such as supervisor behavioral change due to being selected to evaluate, agents responding differently to the evaluator's job assignments, evaluating supervisor\masked supervisors getting better information on CGCS performance, CGCSs trying to influence everyone under masked scheme, etc. We also find suggestive evidence that the improvement in performance is not simply "cheap talk" by supervisors and colleagues: in the masked scheme, the CGCSs are more likely to be assessed as "qualify for tenure," and also receive higher salaries in positions with established incentive pay schemes. We also provide various ways to benchmark the magnitude of our treatment effects.

This paper speaks to three strands of literature. First and foremost, it provides the first rigorous empirical test for the existence and implications of influence activities in workplace. As pointed out by Lazear and Oyer (2012), while a large theoretical literature has studied how agents try to influence the assessments of their supervisors (e.g. Milgrom and Roberts, 1988; Milgrom, 1988; Meyer et al., 1992; Schaefer, 1998), there lacks quality empirical evidence, aside from anecdotes and case studies, to verify these arguments. Our paper fills in this gap by providing field experimental evidence on influence activities among Chinese local state employees, as well as quantifying the causal impacts of alleviating influence activities on job performance. More broadly, subjective performance evaluation is ubiquitous across private and public sectors, and has been investigated extensively by a large body of theoretical works (Gibbons and Murphy, 1992; Baker et al., 1994; Prendergast and Topel, 1996; MacLeod, 2003). However, empirical evidence on the effectiveness and limitations of subjective evaluation is still largely missing, with only a handful of exceptions (Chevalier and Ellison, 1999; Hayes and Schaefer, 2000). Our paper also contributes to this literature by showing how influence activities could undermine the power of subjective performance evaluation.

Second, this paper adds to a growing empirical literature on the personnel economics of the state, specifically on incentivizing state employees (Finan et al., 2016). Most of the existing works on this topic focus on the role of financial incentives,[51] with only a few exceptions studying non-pecuniary incentives, such as transfers and postings (Banerjee et al., 2012), and intrinsic motivation (Ashraf at al., 2014). Our paper adds to this line of work by exploiting the (implicit) career incentive involved in performance evaluations, which is a prevalent form of incentive scheme in public sectors, but has rarely been studied in the literature. In addition, we show that holding the "final reward" fixed, slightly revising the performance evaluation practice can lead to substantial improvement in job performance of state employees, indicating a highly cost-effective way to strengthen state effectiveness.

Third, our paper also relates to the research agenda on Chinese political meritocracy. Since Li and Zhou (2005), a large number of empirical studies have tried to investigate how different performance indicators, such as GDP growth rate (Guo, 2009) and implementation of the one-child policy (Serrato et al., 2018), affect the incentives and behaviors of provincial and prefectural officials in China. However, the existing evidence has been focusing almost exclusively on relatively high-level government officials, leaving the vast majority of grassroots bureaucrats under-researched, whose incentives and constraints could potentially be very different from their high-level leaders. For instance, a key distinction is that, the job tasks for the low-level bureaucrats are much more difficult to be quantified by objective measures such as GDP growth and environmental quality, so most of the grassroots bureaucrats are only rewarded based on subjective evaluations by their supervisors. Our paper intends to unravel the black box of the incentive schemes for grassroots bureaucrats in China, who are the building blocks of state capacity and play key roles in policy implementation and public service delivery.

The remainder of this paper is organized as follows. In Section II, we layout a simple conceptual framework to help guide the experimental design, and generate some testable hypotheses. In Section III, we introduce the institutional background, design, and

[51] See Finan et al. (2016) for a summary.

implementation of our field experiment. In Section IV, we present baseline experimental results to test the main predictions of our conceptual framework. In Section V, we discuss the potential alternative interpretations of the baseline findings. In Section VI, we benchmark the magnitude of our treatment effects. Section VII concludes.

3.2 Conceptual Framework

Assume an agent i works on a job with a productive dimension x_i, which can be observed by co-workers and supervisors, but cannot be verified quantitatively. The organization therefore relies on a subjective performance evaluation scheme, where the reward of the agent depends on the evaluation score given by his supervisor j:

$$Y_{ij} = x_i + u_{ij}$$

To mimic our "duality" empirical setting, we assume that there are two supervisors, $j \in \{a, b\}$. In addition to working on the productive dimension of the job, agent i also has the option to exert (supervisor-specific) influence activities u_{ij} to please supervisor j, in order to improve the evaluation score.

The agent maximizes his expected utility, which is a function of both evaluation outcome and work efforts, subject to a time constraint:

$$\max_{x_i, u_{ij}} \sum_{j \in \{a,b\}} s_{ij} \cdot V\left(Y_{ij},\ L_i\right)$$

s.t.

$$f(x_i) + \sum_{j \in \{a,b\}} g(u_{ij}) + L_i = T$$

where s_{ij} is the probability of each supervisor j's assessment being used to determine individual i's reward in the performance evaluation scheme ($\sum_{j \in \{a,b\}} S_{ij} = 1$); L_i is leisure; $V(Y, L)$ is the utility function of the agent, which is increasing and concave in both evaluation score and leisure. In the time constraint, $f(x)$ is a convex and increasing function measuring the time needed to obtain performance x, $g(u)$ is a convex and increasing function measuring the time needed to exert influence activities u, T is the total time budget for an individual.

Other colleagues also observe the performance, but their opinions are not included in the performance evaluation scheme, so the agent is not incentivized to adjust his efforts to improve colleague evaluations. Colleagues therefore receive no influence activities, and base their evaluations solely on the productive dimension:

$$Y_{ic} = x_i$$

Proposition 1: *when one of the two supervisors is randomly chosen to be the evaluator, and his identity is known to the agent, other things being equal, the evaluating supervisor gives higher assessment than the non-evaluating one.*

Without loss of generality, assume: $s_a = 1, s_b = 0$. Solving the model, we can show that:

$$u_{ia} > u_{ib} = 0$$

which leads to:

$$Y_{ia} > Y_{ib}$$

The intuition is that, when the agent knows the identity of the evaluator, he would exert more evaluator-specific efforts (i.e. $u_{ia} > u_{ib}$) that lead to more positive assessment from the evaluating supervisor ($Y_{ia} > Y_{ib}$).

Proposition 2: *when one of the two supervisors is randomly chosen to be the evaluator, but his identity is unknown to the agent, other things being equal, compared to the case where the identity is known, the performance of the agent will improve, as measured by both supervisor assessment and colleague assessment.*

In this masked scheme, $s_a' = s_b' = \frac{1}{2}$. Since the identity of the evaluator is unknown, the agent would reallocate efforts from supervisor-specific influence activities (u_{ij}) toward productive tasks that would be appreciated by both supervisors (x_i). Therefore, compared to the revealed scheme, we can show:

$$u'_{ia} = u'_{ib} < u_{ia};\ \ L'_i > L_i;\ \ x'_i > x_i$$

and:

$$Y'_{ia} + Y'_{ib} > Y_{ia} + Y_{ib}$$
$$Y'_{ic} > Y_{ic}$$

The intuition is that, since the identity of the evaluator is unknown, the marginal cost of improving expected evaluation outcome through influence activities doubles, while the marginal cost of improving evaluation outcome through the productive dimension remains unchanged. Therefore, the agent re-optimizes by switching from exerting influence activities towards being productive for the organization.

Proposition 2.1: *more risk-averse agents are more strongly incentivized when the identity of the evaluating supervisor is unknown.*

When the agent is more risk averse (i.e. $V(Y,L)''$ being pointwise more concave than $V(Y,L)'$), the reallocation from uncertain dimension (u_{ij}) to certain dimensions (x_i and L_i) would be more salient. So we can show:

$$(Y''_{ia} + Y''_{ib}) - (Y_{ia} + Y_{ib}) > (Y'_{ia} + Y'_{ib}) - (Y_{ia} + Y_{ib})$$

Proposition 2.2: *when at least one of the supervisors has imperfect information about job performance, introducing ex post randomization in evaluator identity is less effective as the information gap enlarges.*

Without loss of generality, assume:

$$Y_{ia} = x_i + u_{ia}; \ \ Y_{ib} = \alpha \cdot x_i + u_{ib}$$

where $\alpha < 1$. Solving the model, it is straightforward to show that:

$$\frac{\partial(Y'_{ic} - Y_{ic})}{\partial\alpha} > 0$$

The intuition is that, when $\alpha \to 0$, x_i becomes more and more specific to supervisor a, therefore similar to the role of u_{ia}. As a result, the effect of "reallocating towards commonly appreciated dimension" is attenuated.

Proposition 2.3: *when job tasks involve multiplicity, introducing ex post randomization in evaluator identity is more effective when the two supervisors' subjective weights for different dimensions of performance are more consistent.*

Without loss of generality, we extend the previous setting by allowing for two dimensions of performance: x_{i1}, x_{i2}, for which a supervisor imposes weights $w_{j1}$and w_{j2} respectively ($w_{j1} + w_{j2} = 1$). So the evaluation function becomes:

$$Y_{ij} = w_{j1} \cdot x_{i1} + w_{j2} \cdot x_{i2} + u_{ij}$$

and the time constraint becomes:

$$f(x_{i1}) + f(x_{i2}) + \sum_{j\in\{a,b\}} g(u_{ij}) + L_i = T$$

Solving the model, we prove that:

$$\frac{\partial[(Y'_{ia} + Y'_{ib}) - (Y_{ia} + Y_{ib})]}{\partial|w_{a1} - w_{b1}|} < 0$$

and:

$$\frac{\partial(Y'_{ic} - Y_{ic})}{\partial|w_{a1} - w_{b1}|} < 0$$

The intuition is that when both supervisors have more consistent subjective weights, it is more desirable to reallocate efforts from supervisor-specific influence activities (u_{ij}) to productive dimensions that would be appreciated by both supervisors (x_{i1}, x_{i2}), which improves both average supervisor evaluation and average colleague evaluation.

3.3 Experiment

Institutional Background

Since the early 2000s, the Chinese government has launched several large-scale human capital transfer programs, which altogether hired more than one million college graduates to work in rural areas, in the hope that their modern human capital and independence from local interest groups could help improve state effectiveness at the local levels. For example, in the College Graduate Village Officials (CGVOs) program, college graduates were hired as village officials on a contractual basis, and it is found that they were able to significantly improve policy implementation and reduce leakage in poverty subsidies (He and Wang, 2017).

In this paper, we focus one of such human capital transfer programs, known as the "3+1 Supports" (*San Zhi Yi Fu*) initiative, which was launched in 2006 by the Ministry of Human Resources and Social Security.[52] Several types of civil service positions are offered to the college graduates through the "3+1 Supports" initiative, including township administrative clerics who work on poverty alleviation, agricultural support, water resource management, and social security provision; teachers in township primary schools; and pharmacists and nurses in township clinics. By the end of 2018, more than 350,000 college graduates had been hired to work as grassroots civil servants through this program. We refer to these individuals as "College Graduates Civil Servants" (CGCSs) in this paper.

The CGCSs are recruited from colleges on a yearly basis. Before each job market season in May, the provincial government announces the total number of vacancies on its website and invites college students to apply. In most provinces, the procedure of CGCS recruitment is similar to that of formal (permanent) state employee recruitment: the applicants need to first take a comprehensive written exam, which is similar to the Administrative Aptitude Test and Essay Writing Test used in the National Civil Service Exam; then, the high-score applicants will be interviewed and the top-ranked candidates will be recruited. Some provinces also recruit CGCSs by review their application materials, where priorities are often given to (Communist) Party members, those who have local *Hukou* (household registration status), those who graduated with honors, and those who have more advanced and relevant degrees.

The admission of CGCSs is highly competitive. For examples, in Shandong province, the total number of available positions in 2017 was around 1,500, but the program attracted more than 31,000 applicants (admission rate < 5%); in Guangxi province, the government planned to hire 800 CGCSs in 2017; but the total number of applicants exceeded 13,600 (less than 6% applicants were recruited). In most provinces and in most years, the admission rate for the "3+1 Supports" program is below 10%. Such intense competition ensures the quality of the selected CGCSs.

A CGCS is similar to a formal entry-level township civil servant in many ways. For CGCSs in clerical positions, like most formal rural civil servants, the job tasks tend to be a combination of routine paper works, interactions with villagers, and case-based assignments from supervisors, making job performance hardly quantifiable, and difficult to compare across individuals. For CGCSs in more specialized positions like township clinic nurses or primary school teachers, the job tasks are also similar to their formal state employee colleagues. While certain dimensions of these jobs are better defined than cleric jobs, the grassroots nature of these jobs still make objective performance evaluation largely infeasible, e.g., student score cannot be used to incentivize teachers due to the lack of unified written exams at the primary school level. As a result, the evaluation of a CGCS

[52] Six other major ministries and departments co-sponsored the program, including the Ministry of Education, the Ministry of Finance, the Ministry of Agriculture, the National Health Commission, the State Council Leading Group Office of Poverty Alleviation and Development, and the Communist Youth League Central Committee.

relies almost solely on the evaluating supervisor's subjective assessment, which is also the *status quo* for the vast majority of civil service jobs across the world.

The only major difference between a CGCS position and a formal civil servant position is that the former is contract-based, while the latter is permanent ("tenured"). The majority of CGCSs are eager to be promoted to formal state employees upon finishing the two-year term, which would only be approved by the government subject to satisfactory job performance. As a result, aspiring CGCSs have strong incentives to impress their supervisors and improve the evaluation outcomes. While the strong incentives could potentially lead to higher work efforts, they might also cause influence activities to please the evaluating supervisor in ways that are inconsistent with the government's interest.[53]

A CGCS is supervised by two leaders: a party leader and an administrative leader. This is due to the unique dual-leadership governance structure in China: in principal, the administrative leader is in charge of the daily operation of the government entity, while the party leader oversees the process; the administrative leader often has larger economic power, while the party leader can exert larger political power. These two leaders have the same official ranking, but normally the party leader is perceived to have an edge in authority. This dual-leadership provides *de facto* checks and balances in China's local governance system (Li, 2018) and is adopted by different levels of administrative units (ranging from the central government to the small village committees). In public institutions such as schools, hospitals, and state owned firms, as long as there are more than three Communist Party members among the employees, this dual-leadership rule needs to be enforced. The dual-leadership setting is important for our experiment, as we will randomly choose one from the two to evaluate the performance of a CGCS.

Experimental Design

In collaboration with two provincial governments in China, we randomize two subjective performance evaluation schemes, a "revealed scheme" vs. a "masked scheme," among all their 3785 CGCSs. In both schemes, one of the two supervisors is randomly selected to be the evaluating supervisor for CGCS performance, meaning that his assessment will be given a 100% weight in the evaluation outcome, which in turn affects future career development of the CGCS. For the other randomized non-evaluating supervisor, we also collect his assessment of CGCS performance, but this assessment is given 0% weight in determining the evaluation outcome. In both schemes, we never directly inform a supervisor whether or not he is chosen as the evaluating supervisor, nor do we inform the colleagues of the CGCS.

In the "revealed" scheme, we inform (and only inform) the CGCS about the identity of his evaluating supervisor, at the beginning of the evaluation cycle. This mimics the status quo of CGCS performance evaluation, where the agent is informed about the evaluating supervisor *ex ante*. The difference is that, in the status quo, the evaluator is endogenously chosen from the two supervisors, typically through an opaque process combining supervisor opinions, division of labor between supervisors, and other idiosyncratic factors. In our "revealed" scheme, by randomly selecting the evaluator, we can tease out the

[53] Simple examples of such influence activities include picking up supervisor's kid from school, making coffee for supervisor, doing chores for supervisor, etc.

endogeneity in evaluator selection, and directly pin down the causal impact of knowing the identity of the evaluator. Two thirds of the CGCSs are assigned to this scheme.

In the "masked" scheme, we do not inform the CGCS about the identity of the evaluator until after the end of the evaluation cycle. Compared to the CGCSs in the revealed scheme, those in the masked scheme do not know the identity of their evaluator throughout the year. As a result, it becomes less beneficial to impose supervisor-specific influence activities, which incentivizes the CGCSs to reallocate their efforts toward productive dimensions that would be appreciated by both supervisors. One third of the CGCSs are assigned to this scheme.

We focus primarily on two sets of outcomes: supervisors' assessments and colleagues' assessments of CGCS performance. We define "colleagues" as co-workers in the same office as the CGCS but are not hired through the "3+1 Supports" program. We consider the colleagues' assessments as the ideal performance measure in this context for three reasons. First, the colleagues were randomly chosen from the same office where the CGCSs worked. They worked closely with CGCSs and can thus best observe the CGCSs' performances. Second, there is no obvious conflict of interest between the CGCSs and their colleagues. Unlike CGCSs who work in the office only under 2-year contracts, most colleagues already have tenure and have worked in the office for many years. As a result, the CGCSs and their senior colleagues do not directly complete with each other in career development. Finally, the CGCSs had no incentive to influence their colleagues for evaluation reasons throughout the study. At the baseline, the provincial governments explicitly told the CGCSs that only the evaluating supervisor's opinions would be used for evaluating their final performance. Most CGCSs did not expect that we would even survey their colleagues until the surveyors were sent to their offices at the end line.

The combination of our experimental design and data collection allows us to test the main propositions of our conceptual framework in powerful ways. First, proposition 1 suggests that in the revealed scheme, the randomly chosen evaluator should give more positive assessments of CGCS performance, due to evaluator-specific influence activities imposed by the CGCS. Since the evaluator is randomly selected from the two leaders, we could directly test this prediction by estimating the causal effect of "being evaluator" on "giving more positive assessment." In addition, for each CGCS, we also elicit his colleagues' beliefs regarding which of the two supervisors would give the CGCS more positive assessments. While we did not inform the colleagues about which co-supervisor was chosen to evaluate, to the extent that supervisor-specific influence activities indeed happened, and could at least be partly observed by others throughout the year, the colleagues should be able to correctly predict that the (randomly selected) evaluator would favor the CGCS.

Second, proposition 2 asserts that when the identity of the evaluator is masked, evaluator-specific influence activities would be alleviated, causing the agent to reallocate efforts toward productive dimensions that would be appreciated by both leaders, which results in better overall performance. The first part of this statement could be verified by seeing whether "leader assessment asymmetry" disappears as we switch from the revealed scheme to the masked scheme; the second part of this statement could be verified by testing whether CGCSs in the masked scheme have better performance as measured by average leader assessment and colleague assessment. Furthermore, our conceptual framework predicts that the masked scheme is most effective when the agent is more risk averse, when the two leaders have more aligned preferences, and when the two leaders have smaller information asymmetry. These additional predictions can be tested by exploring the heterogeneous treatment effects of adopting the masked scheme.

Implementation

Our experiment is conducted in collaboration with two of the largest provinces in China, with a combined population of nearly 150 million. Province A is coastal and more developed, while Province B is inland with lower average income. Province A recruits the CGCSs using an automatic scoring system, while Province B recruits through traditional written tests and interviews. Our sample covers all the 3,785 CGCSs employed by these two provinces in China as of September 2017 (cohorts of 2016 and 2017). Our research team was appointed by the two provincial governments as the "third-party evaluator" for the "3+1 Supports" program to help launch new performance evaluation schemes, which the provincial governments officially informed all the CGCSs. This high-level endorsement helped ensure that the vast majority of the CGCSs were well aware of the high stakes involved in the evaluation outcomes under the newly launched evaluation schemes.

The baseline survey was carried out in September 2017, one month after the 2017-cohort CGCSs finished job training and were assigned to their positions. An online survey was sent to all CGCSs via text message. Through the baseline survey, we gathered a rich set of information on the CGCSs, including their general background, work duties, political beliefs, motivations, risk preferences, etc. During the baseline survey, we asked each CGCS to provide the names, positions and contact information of their two co-supervisors, which we cross-validated with the official government records for accuracy. We also asked each CGCS which of the two supervisors he would prefer to have as the evaluator, which helps us understand the perceived positivity from the standpoint of each CGCS.

Immediately after the baseline survey, we randomized all the CGCSs into the two evaluation schemes based on their work units, and informed each CGCS about the evaluation scheme he was assigned to. The randomization was conducted at the work unit level ("*Dan Wei*") instead of the individual level. Different CGCSs who worked in the same unit (i.e., an office or an organization branch led by the same set of supervisors) were assigned to the same scheme. This is per the request of our government partner to ensure that the evaluation results of CGCSs working in the same unit can be fairly compared to each other. In this setting, because 83.9% of the work units only had one CGCS allocated, randomizing at the work unit level did not hurt our statistical power in any substantial way, as compared to randomizing at the individual level.

To ensure the credibility of our intervention, the provincial governments sent a formal letter with official sealed stamps to every CGCS. The letter emphasized the importance of the performance evaluation, confirmed the evaluation scheme (including the name of the relevant supervisor) specific to the CGCS, and the detailed timeline. We reminded the CGCSs about their evaluation schemes in January 2018.

The end-line survey was carried out in June 2018, and consisted of three parts: colleague assessment, supervisor assessment, and self-assessment. When the surveyors visited the office of a CGCS, if there were less than 5 colleagues in the office, all of them were invited to fill in the colleague questionnaire; if there were more than 5 colleagues, the surveyor randomly sampled 5 of them to fill in the colleague questionnaire, using a random number generator.[54] To protect the privacy of the colleagues and encourage truth-telling, colleague questionnaires were strictly anonymous, and CGCSs were not allowed to communicate with the colleagues during the entire process. The CGCS survey was also conducted on-site, but independently from the colleague survey to avoid interference. Leader assessment was completed online, with an individual-specific link for each leader, listing all the CGCSs in his unit.

In the colleague and leader surveys, we ask questions on basic information of the colleague\leader, his interactions and familiarity with the CGCS, the job tasks of the CGCS, and his assessment of the CGCS along various dimensions. Specifically, we asked for an overall assessment of CGCS performance, as well as a "revealed preference" measure asking the colleague\leader whether they think the GCGS deserves a tenured position in this unit upon finishing his 2-year contract.

The end-line CGCS survey followed a similar structure by asking about interactions with leaders\colleagues and self-assessment along multiple dimensions. In addition, we also asked about a series of questions on career plans, salaries, satisfaction with the "3+1 Supports" program, etc.

Balance and Attrition Tests

To ensure the validity of our experiment, we conduct a battery of relevant balance tests: in the revealed scheme, supervisor characteristics should be balanced between the evaluating and non-evaluating supervisors, which supervisor is selected to evaluate should also be orthogonal to both CGCS and colleague characteristics; between the two evaluation schemes, CGCS characteristics, supervisor characteristics, as well as colleague characteristics should all be balanced. We find that these balance tests all go through, suggesting that our randomization was well executed.

Between the baseline and the end-line surveys, we lost 918 CGCSs (24.3%) in the sample. The main reason for the attrition is that the CGCSs or their supervisors changed their positions during our study period. For example, a CGCS could be relocated from one township to another because of changes in policy targets. The supervisors could be promoted, retired, or move to other institutions. Such job changes would break the supervisor-subordinate relationship and thus invalidate the experimental design. Besides, some CGCSs passed the civil service exam or got admitted by graduate schools (participation in the "3+1 Supports" program gives them extra points in the exam). These people tended to quit their current CGCS jobs and left our experiment. To test whether our experiment caused such attrition, we estimate the relationship between attrition status and the treatment status, we find that our treatment does not lead to selective attrition.

[54] If a colleague was not in the office when the surveyor visited, his contact information was collected and he would be surveyed over the phone in the following day. To ensure data accuracy, the team leader randomly called some colleagues in the following days to verify the sampling procedure and the answers collected.

3.4 Main Results

In this section, we present a series of experimental results consistent with our theoretical predictions. Specifically, we show that: (1) when the identity of the evaluator is revealed to the agent, the assessment given by the evaluating leader is higher than that given by the non-evaluating leader, and that this asymmetry in assessments is correctly predicted by the colleagues of the agent; (2) when the identity of the evaluator is masked, the asymmetry in supervisor assessments disappears, and agent performance improves significantly as measured by average colleague assessment and average supervisor assessment; (3) the evaluating supervisor being more positive under the masked scheme indicates increased effort and decreased utility for the agent; (4) masking evaluator identity is more effective when the agent is more risk-averse, when the two supervisors have similar opinions on what constitute good performance, and when there is little information asymmetry between the two supervisors.

These results are compatible with an interpretation of the agent undertaking influence activities (that target the evaluator and are at the expense of the assigned job) in the revealed scheme, and reallocate efforts from influence activities toward productive dimensions in the masked scheme, as formalized in our conceptual framework. The combination of these findings could already rule out some obvious confounding stories, and we will further confront each of those remaining alternative interpretations in Section V.

Evaluating Supervisor More Positive under and only under the Revealed Scheme

In Table 3.1, the outcome variable is "leader 1 score – leader 2 score," measuring the extra positivity of leader 1 than leader 2 towards the same CGCS. The explanatory variable is a dummy variable indicating whether Supervisor 1 was (randomly) chosen to be the evaluating supervisor.

In Columns (1)–(2), we focus on the revealed scheme, in which every CGCS is informed about the identity of his evaluator at the beginning of the evaluation cycle, who is randomly selected from the two co-supervisors. We find that in the revealed scheme, if a leader is chosen as the evaluator at the baseline, he indeed gives more positive assessment at the end line. In Column (2), we include a rich set of control variables in the regression and the estimated coefficient remains highly stable, confirming that the "Supervisor 1 Evaluating" treatment is indeed randomly assigned.

In Columns (3)–(4), we focus on the masked scheme, where the identity of the randomly chosen evaluator is not announced until the end of the evaluation cycle. As we can see, in the masked scheme, being selected as the evaluator no longer has any impact on the positivity of assessment, again confirming our theoretical prediction.

The asymmetry in supervisor assessments under the revealed scheme, and its disappearance under the masked scheme, can be rationalized by the agent imposing supervisor-specific influence activities for better evaluation outcomes when and only when he knows the target.

Evaluation Asymmetry Predicted by Colleagues under and only under the Revealed Scheme

If the results in Table 3.1 are indeed driven by influence activities, to the extent that such behaviors can at least be partially observed by other co-workers in the same office, we should expect that in the revealed scheme, colleagues could update their priors on which of the two leaders would be more positive about CGCS performance. In other words, when colleagues receive noisy signals of u_{ij}, even without knowing who is the evaluator, they would be more likely to correctly identify the more positive leader. We test this prediction in Table 3.2, where the outcome of interest is a dummy variable indicating "whether a colleague thinks leader 1 would be more positive."

Columns (1)–(2) look at the revealed scheme. We see that when leader 1 is randomly selected as the evaluator, colleagues are more likely to think he is going to give more positive assessments. Combined with the results in Table 3.1, colleagues indeed correctly predict the direction of leader assessment asymmetry.

In Columns (3)–(4), we investigate the masked scheme. As can be seen, when the identity of the evaluating leader is unknown to the agent, colleagues are also no longer more likely to identify the evaluator as the more positive leader.

Combined together, these results are consistent with a case where agents impose evaluator-specific influence activities when knowing who the evaluator is, and such behaviors could be observed by other colleagues who work in the same unit and share the same office.

Improved Performance under Masked Scheme

We exploit the random assignment of CGCSs between the two evaluation schemes to estimate the causal effect of masking evaluator identity on job performance.

First, in Table 3.3, we examine colleague assessments, which we consider as the ideal measure of CGCS performance, for reasons explained in Section III. Specifically, the dependent variable is average colleague assessment of the CGCS's performance, which is framed relative to other civil servants working in the same work unit. The assessment score ranges from 1 to 7, representing different categories from "worse than all the colleagues" to "better than all the colleagues" in the questionnaire. Relatedly, we have an outcome variable indicating whether the colleagues think the CGCS's performance ranks in the Top 10% of the organization.

The results in Columns (1)–(2) show that masking the identity of the evaluator significantly improved colleague assessments of CGCS performance. In Columns (3)–(4), we use the "Top 10%" as the outcome variable, all the results remain similar: CGCSs in the second treatment arm are more likely to be recognized as top performers by their colleagues.

These results are consistent our theoretical prediction that masking evaluator identity would reduce influence activities, thus improving job performance. To corroborate this finding and better understand the driving-force behind performance improvement, we also investigate the impacts on supervisor assessment.

Table 3.4 reports the treatment effects (of being assigned to the masked scheme rather than the revealed scheme) on supervisor assessments. In Columns (1), the outcome variable is the mean assessment of the two supervisors. We find that the "masking the identity of the evaluator" significantly improves average supervisor assessment. To understand how the CGCSs manage to achieve overall higher assessments, we look at the treatment effects on evaluator leader and non-evaluating leader separately. Columns (2) and (3) reveal that both the evaluating and non-evaluating supervisor's assessment improved under the masked scheme, with a larger improvement in the non-evaluating supervisor's assessment. The difference between evaluating and non-evaluating leaders' treatment effects motivate Column (4): deviation between leader assessments is smaller under the masked scheme.

Rationalizing Results on Supervisor Evaluation

In Table 3.4, the result for the non-evaluating supervisor's assessment is intuitive: the CGCS now needs to impress both supervisors, instead of one, making the non-evaluating supervisor more pleased. The result for the evaluating supervisor's assessment, however, is less intuitive: the evaluating supervisor's assessment score is also improved, despite no longer receiving influence activities catered to his/her private utility. This result is rationalized in our model: if in scheme 2, the agent sacrifices enough of his leisure to work on productive dimensions, this "labor participation effect" could dominate the "reduced influence activity" effect, leading to an increase in the evaluation from the more favorable supervisor.

We test this intuition from our model in Table 3.5. In our survey, besides asking the colleagues to give an overall assessment of the CGCS's performance, we also asked them to specifically tell us whether they think the CGCS work harder and overtime relative to other colleagues. In Panel A of Table 3.5, we find that masking the evaluator's identity indeed makes the CGCS to work harder and overtime (as observed by their colleagues), suggesting the CGCS allocate more efforts in the productive dimensions.

In addition, we report the findings for the CGCSs' self-evaluations. Note that results from self-evaluations are less straightforward to interpret, because the incentives behind the self-reporting are complicated. A vast psychology literature has looked at this issue but the validity of self-evaluation remains controversial.[55] In our setting, for example, almost all of the CGCSs thought that they performed better than their colleagues and the majority of them were in the Top 10% category. They generally exaggerated their work performances, which resulted in smaller variations in self-evaluated performance scores. Acknowledging various problems of self-evaluated outcomes, here, however, we still find that the "masking the identity of the evaluator" positively affected the CGCSs' self-evaluations.

An important prediction of our model is that under the uncertainty treatment, as the CGCS needs to allocate more efforts to the productive dimension, his overall welfare would be decreased. In our survey, we asked the CGCS whether they would be willing to find a job in the private sector, and if so, what the expected wage is. In Panel B of Table 3.5, we find that masking the identity of the evaluator makes the CGCS more willing to accept a job in the private sector, and the "reserve wage" is reduced by about 5%. Our

[55] See Hetts (1999) for a review of the psychology literature.

interpretation is that because with uncertainty they had to work harder, and it made them more willing to quit and accept a low-pay job. These results help rationalize our previous finding that both the evaluating and non-evaluating supervisors' assessment improved.

Risk Aversion, Information Gap, Weights on Tasks, and Supervisor Importance

In Table 3.6, we explore heterogeneous treatment effects under the guidance of our theoretical predictions. Combined together, these results shed light on the channels through which "masking evaluator identity" improves CGCS performance.

In the conceptual framework, we predict that more risk-averse agents will respond more strongly to the introduction of uncertainty. This is confirmed by the results in Column (1), where the interaction term of risk averse dummy and the introduction of the masked scheme is positive and significant.[56]

Our model also predicts that when the information gap of the two supervisors gets larger, the "productive dimension" effectively becomes more and more supervisor-specific. As a result, larger information asymmetry between the two supervisors weakens the treatment effect of masking evaluator identity. This is confirmed by both Column (2), where we use "difference in two supervisors' self-reported information" to measure information asymmetry, and Column (4), where we use "two supervisors' difference in work assignment frequency" to proxy for information asymmetry.

Another prediction of the model is that when the preferences of the two supervisors are highly aligned, masking evaluator identity should be most powerful. In our survey, we separately elicited each supervisor's subjective ranking on the importance of each performance dimension. In Column (3), we find that the treatment effects are larger when the two supervisors give the exact same ranking of importance, confirming the prediction. This result suggests that the treatment effect of masking evaluator identity is not driven by "CGCS responding to more diversified leader preferences," because if that is the case, we should see a weaker treatment effect when leaders have more aligned preferences, not stronger.

3.5 Alternative Interpretations of Baseline Results

Our model suggests that under the revealed scheme, an agent imposes evaluator-specific influence activities; and under the masked scheme, the agent no longer knows who is evaluating, so would re-optimize his efforts by reducing influence activities to work on productive dimensions that would be appreciated by both leaders. In Section IV, we documented a series of evidence supporting the main predictions of our model.

While some confounding interpretation might be consistent with one of the baseline findings, it could hardly rationalize the combination of all these patterns at the same time. Nevertheless, in this section, we treat each of the baseline result separately, and examine the alternative explanations specific to it.

[56] Risk attitude is elicited through a hypothetical coin-flipping game. An individual taking a certain 400 Yuan over a 50% chance to win 2000 Yuan is defined as highly risk averse. Our results are robust to alternative cutoffs for this definition.

A. Confounders for "Leader Assessment Asymmetry"

Confounder 1: Behavioral Differences between the Evaluating and Non-Evaluating Supervisors

Our interpretation of the findings in Table 3.1 is that the subordinate is able to impose supervisor-specific influence activities, and such behaviors can be observed by their co-workers. An alternative explanation is that the evaluating and non-evaluating supervisors may act differently, simply because that "revealing the identity of the evaluator" might directly affect the evaluator's behaviors. For example, if the evaluating supervisor feels more responsible after knowing that he is the evaluator, he may give the CGCS more positive assessments to avoid potential psychological burdens. Although this concern is alleviated by our experimental design that the supervisors were not informed by the research team about their roles in the evaluation, it is still possible that the CGCSs might delivered this information to them.

To investigate this probability, in our end-line survey, we directly asked the supervisors about whether they were aware of their responsibility in evaluating the CGCS. It turns out that the majority of them (more than 65%) did not know that they were chosen as the evaluators until after the end of the evaluation cycle. In Panel A of Table 3.7, we estimate the impact of "Supervisor 1 Evaluating" on "Supervisor 1 Being More Positive" separately for different subsamples: the subsample in which the supervisors did not know their (evaluator) roles, and the subsample in which the supervisors knew their roles (likely through the CGCS). We find that in both subsamples, "Supervisor 1 Evaluating" increases the likelihood that "Supervisor 1 Being More Positive," suggesting that our results are not driven by supervisor behavioral change due to being evaluator.

Moreover, in Panel B of Table 3.7, we directly examine the existence of behavioral differences between the evaluating and non-evaluating supervisors. We focus on three outcomes: (1) the likelihood of a supervisor not responding the survey, (2) whether one supervisor writes more words in describing the subordinated CGCS's talks than the other supervisor, and (3) whether one supervisor assigns more job tasks to the CGCS than the other. Our hypothesis is that, if the evaluating supervisor indeed paid more attention to the CGCS, we should observe the evaluating supervisor to be more likely to answer the survey, write more words in their assessment, and even assign more tasks to the CGCS. Again, our analyses do not support this hypothesis. "Supervisor 1 Evaluating" does not affect any of these outcomes, as evidenced by the regression results in Panel B. This is again inconsistent with the "behavioral differences" interpretation.

Confounder 2: Higher Information Quality for Evaluating-Supervisor

Another confounding story is that even without behavioral changes, the evaluating supervisor would receive more information regarding CGCS performance from various sources: the CGCS, the colleagues, and the other (non-evaluating) supervisor might try to send signals to help him better evaluate. This improvement in information quality might improve supervisor assessment, while the CGCS did not actually exhibit any "influence activities" at the cost of the organization.

To examine this interpretation, in our end-line supervisor survey, we asked each supervisor how frequent did the subordinated CGCS, the colleagues of the CGCS, or the other supervisor discuss with him about CGCS performance. We are interested in whether the evaluating supervisor would receive additional information than the non-evaluating supervisor from these sources. In panel C of Table 3.7, we find that "Supervisor 1 Evaluating" does not make supervisor 1 obtain extra information from any of these sources, as compared to supervisor 2. In other words, revealing the identity of the evaluator to the CGCS does not differentially change the "information quality" of the evaluating supervisor.

B. Confounders for "Treatment Effects of Masking Evaluator Identity"

Confounder 1: CGCS Influencing Everyone under Masked Scheme

We find that CGCSs in the masked scheme on average receive more positive assessments from colleagues and both leaders, which we interpret as indicating better job performance.

A potential alternative interpretation is that under the masked scheme, CGCSs do not work harder on the productive dimensions. Instead, they simply try to "influence" everyone, i.e., providing personal favors to all their colleagues and leaders, which leads to more positive assessments from everyone. This is unlikely to be the case, because the CGCSs have no incentive to do so: we clearly indicated at the beginning of the experiment that only the evaluating leader's assessment will determine the evaluation outcome. Nevertheless, we try to examine this possibility more formally.

If the CGCSs are indeed trying to influence their colleagues and leaders more often, we would expect that they interact more with these individuals, and the colleagues and leaders become more familiar with the CGCSs, in both work and life. However, in Table 3.8 Panel A, we find that "masking the identity of the evaluator" did not make the CGCSs to communicate more or meet more frequently with their colleagues (Columns (1) and (2)), nor did it make the colleagues become more familiar with the CGCS's work or life (Columns (3) and (4)). This is again inconsistent with the alternative interpretation.

Confounder 2: Higher Information Quality in Masked Scheme

Another possibility is that the leaders in the masked scheme tend to get better information on CGCS performance, which might explain the improved leader assessments.

To address this concern, in Panel B of Table 3.8, we examine whether leaders get additional information on CGCS performance under the masked scheme, either from colleagues or the CGCS himself. We find that, for both the evaluating leader and the non-evaluating leader, being in the masked scheme does not increase the frequency of colleagues and CGCSs reporting to them regarding CGCS performance. This suggests that the improved supervisor evaluations in the masked scheme cannot be explained by changes in information quality.

3.6 Benchmarking Effect Size

Usually, subjective evaluations are adopted when there lack clearly objective performance measures. Due to this nature, it is often challenging to benchmark the improvement in subjective evaluations using objective indicators. To better interpret the economic significance of our "masked scheme" intervention, we use two different ways to provide some more intuitive benchmarks for the magnitude of our estimated coefficients.

Comparing Magnitudes with Correlative Coefficients

In Table 3.9, we report the correlations between subjective evaluations and a rich set of CGCSs' characteristics. In Column (1), we see that age (or more experience), education and party member have the strongest predictive power in explaining colleagues' assessments. Being one year older is associated with a 0.05-point increase in colleague assessment, graduating from a 4-year college instead of a 3-year community college is correlated with a 0.15-point increase in colleague assessment, and being a party member is associated with 1 0.17-point increase in colleague assessment. In Column (2), we estimate the correlations between supervisors' evaluations and CGCSs' personal characteristics. We find similar patterns that CGCSs' age, education and party membership are positively associated with better evaluation outcomes. The magnitudes of the coefficients are similar to those in Column (1).

Recall in Table 3.3, "masking the identity of the evaluator" can increase colleagues' assessments by 0.22. The treatment effect is almost "equivalent" to having 4 more years of (life and work) experience, and is substantially larger than the effect of "upgrading" all 3-year college graduates to 4-year college graduates, or replacing all non-CCP members with CCP-members. Given the substantial edge in prestige and ability associated with 4-year colleges and CCP member status, the treatment effect does seem to be economically significant.

"Revealed Preference" Measures

The main performance measures used in this paper are subjective assessments given by different parties. This might raise concerns about the economic significance of our findings, if there is any reason that the incentives to "cheap talk" would vary systematically based on our random assignment of evaluation schemes. To address this potential issue, in Table 3.10, we examine two "revealed preference" outcomes, which involve real stakes and cannot be explained by cheap talk.

First, we directly ask the colleagues whether they think the CGCSs should be offered a tenured position in the office after the 2-year term. The question was posited at the end of the survey, and many of the colleagues asked our surveyors whether their answers to this question would be referenced by upper level government in the decision-making process. While we told them that all the information in the survey would be useful, their particular cautiousness on this question suggests that they were more aware of the stakes involved. Using this as an alternative outcome, we find that masking the evaluator's identity indeed leads more colleagues to think the CGCS should get tenured, as reported in Column (1) of Table 3.10.

Second, we asked each CGCS to report their total monthly remuneration, which includes basic wages and bonuses (if there is any), which we in turn verified using administrative information provided by the provincial government. The basic wage is set by upper-level government and should be exactly the same for all the CGCSs, conditional on their county of residence, enrollment year, and positions. Each working unit in turn has some discretion to reward the best performing employees with a modest amount of bonus (usually very small in the government system). In Columns (2) to (3) of Table 3.10, we observe that the CGCSs in the masked scheme earn 2% (roughly 49 RMB per month, 0.3 S.D.) higher salaries than those in the revealed scheme. Since the base salary for CGCSs is fixed (matched to the entry-level permanent civil servant wage), this 2% gap completely reflects the difference in bonuses.

During our field interviews, we were informed that mainly those CGCSs who work as nurses enjoy bonuses, which is a function of objective performance indicators such as number of patients taken care of and number of night shifts completed. Indeed, when we restrict the sample to nurses, the coefficient gets substantially larger (116 RMB). This suggests that the improved performance is not cheap talk, instead, it could be reflected in objective productivity measures, and is attached to real stakes.

3.7 Conclusion

Subjective evaluations are widely used in both public and private sectors, especially in contexts lacking sharp measures of employee efforts and performances. A key limitation of subjective evaluation is that it may distort employees' incentives and make them more likely to cater to the preferences of the evaluator, rather than focusing on productive tasks that benefit the organization. However, rigorous empirical evidence on the existence of influence activities is rare and the implications of such activities remain under-explored.

To shed light on these issues, we conduct a large-scale field experiment, where we randomize two performance evaluation schemes among 3,785 junior state employees in China. In the first scheme, we randomly choose one supervisor (from two) as the performance evaluator, and inform the subordinate about the identity of the evaluator *ex ante*. We find that doing so indeed increases evaluator-specific influence activities, as the evaluating supervisor on average gives more positive assessment of the subordinate's performance than the non-evaluating supervisor. Furthermore, other colleagues could correctly predict that the randomly selected evaluator would favor the subordinate (as compared to the non-evaluating supervisor), which indicates that the subordinate's supervisor-specific influence activities could be partly observed by their fellow co-workers.

In the second intervention, the identity of the evaluator is not disclosed to the subordinate, which in theory should encourage the subordinate to reallocate supervisor-specific efforts toward productive dimensions that could be appreciated by both supervisors. We find that this intervention indeed improves the work performance of the subordinate, as measured both by both colleague and supervisor assessments. These results provide causal evidence that reducing influence activities leads to better work performance in the government.

In addition to providing the first rigorous empirical evidence on the existence and implications of influence activities, our findings also have important policy implications. We find that by randomizing the identity of the evaluating supervisor, which has minimal implementation cost, the government could significantly improve the job performance of its employees. This intervention could be adopted to incentivize the 39.5 million state employees in China, and might also apply to broader labor markets given the prevalence of influence activities in workplaces across countries and sectors.

Table 3.1 Revealing Evaluator Identity Causes Scoring Asymmetry

	Supervisor 1 Score - Supervisor 2 Score			
	(1)	(2)	(3)	(4)
Supervisor 1 Evaluating (Revealed)	0.310***	0.310***		
	(0.082)	(0.082)		
Supervisor 1 Evaluating (Masked)			0.003	0.033
			(0.113)	(0.114)
Outcome Mean	-0.019	-0.019	-0.005	-0.005
Outcome S.D.	1.299	1.299	1.217	1.217
Controls	N	Y	N	Y
County FE	Y	Y	Y	Y
Type FE	Y	Y	Y	Y
Enrol Year FE	Y	Y	Y	Y
Observations	1,301	1,301	580	580
R-squared	0.160	0.162	0.242	0.254

Notes: Each column represents a separate OLS regression. Standard errors clustered at work unit level are reported below the coefficients. Control variables include CGCSs' personal characteristics. * significant at 10% ** significant at 5% *** significant at 1%.

Table 3.2 Colleagues (Correctly) Speculate that the Evaluating Supervisor is More Positive

	Colleagues Predict Supervisor 1 Being More Positive			
	(1)	(2)	(3)	(4)
Supervisor 1 Evaluating (Revealed)	0.058***	0.057***		
	(0.020)	(0.020)		
Supervisor1 Evaluating (Masked)			0.008	0.009
			(0.019)	(0.019)
Outcome Mean	0.555	0.555	0.546	0.546
Outcome S.D.	0.497	0.497	0.498	0.498
Controls	N	Y	N	Y
County FE	Y	Y	Y	Y
Type FE	Y	Y	Y	Y
Enrol Year FE	Y	Y	Y	Y
Observations	5,582	5,582	2,615	2,615
R-squared	0.160	0.162	0.242	0.254

Notes: Each column represents a separate OLS regression. Standard errors clustered at work unit level are reported below the coefficients. Control variables include colleagues' personal characteristics. * significant at 10% ** significant at 5% *** significant at 1%.

Table 3.3 Treatment Effects on Colleague Evaluations

	Performance (1-7)		Top 10%	
	(1)	(2)	(3)	(4)
Masking Evaluator's Identity (T2)	0.220***	0.185***	0.081***	0.070***
	(0.033)	(0.027)	(0.012)	(0.011)
Colleague Controls	N	Y	N	Y
County FE	Y	Y	Y	Y
Type FE	Y	Y	Y	Y
Enrol Year FE	Y	Y	Y	Y
Obs.	9,256	9,167	9,256	9,167
R-Squared	0.130	0.371	0.084	0.267

Notes: Each column represents a separate OLS regression. Standard errors clustered at work unit level are reported below the coefficients. Control variables include colleagues' personal characteristics. * significant at 10% ** significant at 5% *** significant at 1%.

Table 3.4 Treatment Effects on Supervisor Evaluations

	Mean (Superior Evaluations)	Evaluating Supervisors' Evaluation	Non-Evaluating Supervisors' Evaluation	Supervisor Deviation
	(1)	(2)	(3)	(4)
Mask	0.138***	0.095*	0.184***	-0.100**
	(0.046)	(0.056)	(0.058)	(0.050)
Controls	Y	Y	Y	Y
County FE	Y	Y	Y	Y
Type FE	Y	Y	Y	Y
Cohort FE	Y	Y	Y	Y
Obs.	1,945	1,940	1,940	1,945
R-Squared	0.255	0.210	0.204	0.132

Table 3.5 CGCSs Work Harder and Become More Willing to Leave

	(1)	(2)	(3)	(4)
Panel A. CGCSs Work Harder				
	Work Hard and Overtime According to Colleagues' Observation		Self Evaluation (1-7)	
Masking Evaluator's Identity (T2)	0.023**	0.023**	0.081*	0.080*
	(0.012)	(0.012)	(0.048)	(0.048)
Obs.	9,349	9,349	2,771	2,771
R-Squared	0.491	0.491	0.117	0.125
Panel B. CGCSs' Welfare Revealed by Reserved Wage				
	Willingness to Accept Private Sector Job		Reserved Wage (log)	
Masking Evaluator's Identity (T2)	0.028*	0.028*	-0.047*	-0.047*
	(0.016)	(0.016)	(0.026)	(0.026)
Obs.	2,735	2,735	2,736	2,736
R-Squared	0.234	0.235	0.233	0.233
Controls	N	Y	N	Y
County FE	Y	Y	Y	Y
Type FE	Y	Y	Y	Y
Enrol Year FE	Y	Y	Y	Y

Notes: Each column represents a separate OLS regression. Standard errors clustered at work unit level are reported below the coefficients. * significant at 10% ** significant at 5% *** significant at 1%.

Table 3.6 Mechanism: Risk Aversion and Relative Importance of Supervisors

	Colleague Evaluation (1-7)			
	(1)	(2)	(3)	(4)
Risk Aversion	-0.044			
	(0.031)			
Masking*Risk Aversion	0.105**			
	(0.053)			
Supervisors' Info. Gap		0.195*		
		(0.101)		
Masking*Supervisors' Info. Gap		-0.365*		
		(0.189)		
Supervisors' Weights Similarity			-0.088	
			(0.069)	
Masking*Weights Similarity			0.212*	
			(0.112)	
Δ in Superiors' Work Assign. Freq.				0.003***
				(0.001)
Masking*Δ in Work Assign. Freq.				-0.002***
				(0.001)
Masking Evaluator's Identity (T2)	0.140***	0.526***	0.173***	0.208***
	(0.037)	(0.176)	(0.030)	(0.029)
Controls	Y	Y	Y	Y
County FE	Y	Y	Y	Y
Type FE	Y	Y	Y	Y
Enroll Year FE	Y	Y	Y	Y
Obs.	9,221	8,788	8,770	9,243
R-Squared	0.355	0.361	0.397	0.335

Notes: Each column represents a separate OLS regression. Control variables include colleagues' characteristics. Standard errors clustered at work unit level are reported below the coefficients. * significant at 10% ** significant at 5% *** significant at 1%.

Table 3.7 Ruling Out Alternative Explanations to Influence Activities

	(1)	(2)	(3)
Panel A. Does Supervisor Evaluation Depend on their Awareness?			
	Supervisor 1 is More Positive		
Supervisor 1 Eva. (*ex ante*)	0.068***	0.063**	0.075
	(0.024)	(0.029)	(0.051)
Sample	Full Sample	Supervisor 1 Unaware of being the Evaluator	Supervisor 1 Aware of Being the Evaluator
Obs.	1,502	1,046	380
R-Squared	0.140	0.172	0.281
Panel B. Behavioral Changes of the Evaluating Supervisor?			
	Sup 1 Not Responding to the Survey	Sup.1 Writes More Words in Describing CGCS's Job	Sup. 1 Assigns More Tasks to the CGCS
Supervisor 1 Eva. (*ex ante*)	-0.010	0.649	0.236
	(0.019)	(0.431)	(0.181)
Obs.	1,910	1,910	1,910
R-Squared	0.144	0.147	0.144
Panel C. Does the Evaluating Supervisor Receive More Information?			
	Supe 1 Gets More Information from CGCS than Sup 2	Sup 1 Gets More Information from Colleagues than Sup 2	Sup 1 Gets More Info from other Sup
Supervisor 1 Eva. (*ex ante*)	-0.012	-0.010	0.017
	(0.016)	(0.017)	(0.020)
Obs.	1,910	1,910	1,910
R-Squared	0.123	0.146	0.157
County FE	Y	Y	Y
Type FE	Y	Y	Y
Enrol Year FE	Y	Y	Y

Notes: Each column represents a separate OLS regression. Standard errors clustered at work unit level are reported below the coefficients. In Columns (1) and (2), control variables include CGCSs' personal characteristics. In Columns (3) and (4), control variables include colleagues' personal characteristics. * significant at 10% ** significant at 5% *** significant at 1%.

Table 3.8 Alternative Explanations to the Treatment Effects of Uncertainty

	(1)	(2)	(3)	(4)
Panel A. Do CGCGs Influence All Their Colleagues?				
	Communication with Colleagues	Meeting with Colleagues	Colleagues Familiar with CGCS Work	Colleagues Familiar with CGCS Life
Masking Evaluator's Identity (T2)	-0.008	0.013	0.020	0.066
	(0.013)	(0.020)	(0.034)	(0.059)
Obs.	9,272	9,349	9,252	9,244
R-Squared	0.055	0.066	0.066	0.083
Panel B. Masking the Identity of the Supervisor Does Not Lead to Information Difference				
	Information Diff. between the Evaluating Supervisors in T1 and T2		Information Diff. between the Non-Evaluating Supervisors in T1 and T2	
Masking Evaluator's Identity (T2)	0.022	0.016	-0.007	-0.001
	(0.016)	(0.017)	(0.016)	(0.017)
Information Provided by	CGCSs	Colleagues	CGCSs	Colleagues
Obs.	2,839	2,839	2,839	2,839
R-Squared	0.125	0.109	0.114	0.113
County FE	Y	Y	Y	Y
Type FE	Y	Y	Y	Y
Enrol Year FE	Y	Y	Y	Y

Notes: Each column represents a separate OLS regression. Standard errors clustered at work unit level are reported below the coefficients. * significant at 10% ** significant at 5% *** significant at 1%.

Table 3.9 Correlations between CGCS Characteristics and Colleague Evaluation

	Performance (1-7)	
	Colleague	Supervisor
	(1)	(2)
Age	0.049***	0.058***
	(0.011)	(0.015)
Gender	-0.069*	-0.055
	(0.038)	(0.040)
Social Science	-0.005	-0.004
	(0.038)	(0.040)
4-Year College	0.153***	0.129***
	(0.048)	(0.048)
STEM Students	-0.065	-0.009
	(0.041)	(0.042)
Party Member	0.173***	0.143***
	(0.041)	(0.051)
Parent High Sch.	0.033	0.044
	(0.043)	(0.047)
Parent College	-0.030	0.096*
	(0.047)	(0.051)
Work in Village	0.004	0.090*
	(0.060)	(0.051)
CEE Score	-0.038	0.005
	(0.032)	(0.035)
Risk Averse	-0.009	-0.009
	(0.030)	(0.040)
Obs.	8,871	2,556
R-Squared	0.017	0.032

Notes: Each column represents a separate regression. Column (1) reports the correlation between CGCSs performances evaluated by their colleagues and the CGCSs' personal characteristics. Column (2) reports the correlation between CGCSs' performances evaluated by their supervisors and the CGCSs' personal characteristics. Standard errors clustered at work unit level are reported in the parentheses. * significant at 10% ** significant at 5% *** significant at 1%.

Table 3.10 Treatment Effects on "Revealed Preference" Measures

	Qualify for Tenure	log(Wage)	Wage	Wage (Medical Support)
	(1)	(2)	(3)	(4)
Masking Evaluator's Identity (T2)	0.032***	0.020**	48.812**	115.537*
	(0.009)	(0.008)	(22.411)	(61.936)
County FE	Y	Y	Y	Y
Type FE	Y	Y	Y	Y
Enroll Year FE	Y	Y	Y	Y
Obs.	9,349	2,750	2,750	193
R-Square	0.099	0.665	0.641	0.737

Notes: Each column represents a separate OLS regression. Control variables include CGCSs' personal characteristics. Standard errors clustered at work unit level are reported below the coefficients. * significant at 10% ** significant at 5% *** significant at 1%.

References

Acemoglu, Daron, and James A. Robinson. "Economics versus politics: Pitfalls of policy advice." Journal of Economic Perspectives 27, no. 2 (2013): 173-92.

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer. 2015. Identification Properties of Recent Production Function Estimators. Econometrica, 83(6), pp.2411-2451.

Alder, Simon, Lin Shou, and Fabrizio Zilibotti. "Economic reforms and industrial policy in a panel of Chinese cities." (2013).

Ambec, Stefan, Mark A. Cohen, Stewart Elgie, and Paul Lanoie. 2013. "The Porter Hypothesis at 20: Can Environmental Regulation Enhance Innovation and Competitiveness?" Review of Environmental Economics and Policy 7 (1): 2-22.

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. "No margin, no mission? A field experiment on incentives for public service delivery." Journal of Public Economics 120 (2014): 1-17.

Baker, George, Robert Gibbons, and Kevin J. Murphy. "Subjective performance measures in optimal incentive contracts." The Quarterly Journal of Economics 109, no. 4 (1994): 1125-1156.

Banerjee, Abhijit V., Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. "Can institutions be reformed from within? Evidence from a randomized experiment with the Rajasthan police." (2012).

Banzhaf, H. Spencer, and Randall P. Walsh. "Do people vote with their feet? An empirical test of Tiebout's mechanism." The American Economic Review (2008): 843-863.

Becker, Randy, and Vernon Henderson. 2000. "Effects of Air Quality Regulations on Polluting Industries." Journal of Political Economy 108 (2): 379-421.

Berman, Eli, and Linda TM Bui. 2001. "Environmental Regulation and Productivity: Evidence from Oil Refineries." Review of Economics and Statistics 83 (3): 498-510.

Boyenge, Singa, Jean-Pierre. "ILO database on export processing zones." International Labour Organization, 2007.

Brennan, Geoffrey, and James M. Buchanan. "The power to tax: Analytic foundations of a fiscal constitution." Cambridge University Press, 1980.

Brueckner, Jan K. "A test for allocative efficiency in the local public sector." Journal of Public Economics 19.3 (1982): 311-331.

Brueckner, Jan K. "Strategic interaction among governments: An overview of empirical studies." International Regional Science Review 26.2 (2003): 175-188.

Bucovetsky, Sam. "Asymmetric tax competition." Journal of Urban Economics 30.2 (1991): 167-181.

Burgess, R., Hansen, M., Olken, B.A., Potapov, P. and Sieber, S., 2012. "The political economy of deforestation in the tropics." The Quarterly Journal of Economics, 127(4), pp.1707-1754.

Cai, Xiqian, Yi Lu, Mingqin Wu, and Linhui Yu. 2016. "Does Environmental Regulation Drive Away Inbound Foreign Direct Investment? Evidence from a Quasi-Natural Experiment in China." Journal of Development Economics 123: 73-85.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. Forthcoming. "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference.” Journal of the American Statistical Association.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. “Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs.” Econometrica 82 (6): 2295-326.

Cameron, A.C. and Miller, D.L., 2015. “A practitioner’s guide to cluster-robust inference.” Journal of Human Resources, 50(2), pp.317-372.

Cattaneo, M. D., Michael Jansson, and Xinwei Ma. 2017a. “Simple Local Polynomial Density Estimators.” Working paper, University of Michigan.

Cattaneo, M. D., Michael Jansson, and Xinwei Ma. 2017b. “rddensity: Manipulation Testing based on Density Discontinuity.” Stata Journal, forthcoming.

Chen, Shuo., and Yiran Li. “Secrets of ghost towns: Career incentives and politically driven urbanization in China.” (2015).

Cheng, Ming-Yen, Jianqing Fan, and James S. Marron. 1997. “On Automatic Boundary Corrections.” The Annals of Statistics 25 (4): 1691-708.

Cheng, Yiwen. “Place-based policies in a development context: Evidence from China.” (2015).

Deb, Joyee, Jin Li, and Arijit Mukherjee. "Relational contracts with subjective peer evaluations." The RAND Journal of Economics 47, no. 1 (2016): 3-28.

Ebenstein, Avraham. 2012. “The Consequences of Industrialization: Evidence from Water Pollution and Digestive Cancers in China.” Review of Economics and Statistics 94 (1): 186-201.

Ebenstein, Avraham, Maoyong Fan, Michael Greenstone, Guojun He, and Maigeng Zhou. "New evidence on the impact of sustained exposure to air pollution on life expectancy from China’s Huai River Policy." Proceedings of the National Academy of Sciences 114, no. 39 (2017): 10384-10389.

Epple, Dennis., and Holger Sieg. "Estimating equilibrium models of local jurisdictions." Journal of Political Economy 107.4 (1999): 645-681.

Fan, Shenggen, Lixing Li, and Xiaobo Zhang. "Challenges of creating cities in China: Lessons from a short-lived county-to-city upgrading policy." Journal of Comparative Economics 40.3 (2012): 476-491.

Farole, Thomas, and Gokhan Akinci, eds. “Special economic zones: progress, emerging challenges, and future directions.” World Bank Publications, 2011.

Finan, Frederico, Benjamin A. Olken, and Rohini Pande. The personnel economics of the state. No. w21825. National Bureau of Economic Research, 2015.

Fredriksson, Per G., John A. List, and Daniel L. Millimet. 2003. “Bureaucratic Corruption, Environmental Policy and Inbound US FDI: Theory and Evidence.” Journal of Public Economics 87 (7): 1407-30.

Fredriksson, Per G., and Daniel L. Millimet. "Strategic interaction and the determination of environmental policy across US states." Journal of Urban Economics 51.1 (2002): 101-122.

Gelman, Andrew, and Guido Imbens. 2017. “Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs.” Journal of Business and Economic

Statistics.

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.

Gibbons, Robert, and Kevin J. Murphy. "Optimal incentive contracts in the presence of career concerns: Theory and evidence." Journal of political Economy 100, no. 3 (1992): 468-505.

Greenstone, Michael. 2002. "The Impacts of Environmental Regulations on Industrial Activity: Evidence from the 1970 and 1977 Clean Air Act Amendments and the Census of Manufactures." Journal of Political Economy 110 (6): 1175-219.

Guo, Gang. "China's local political budget cycles." American Journal of Political Science 53, no. 3 (2009): 621-632.

Cameron, A. Colin, and Douglas L. Miller. "A practitioner's guide to cluster-robust inference." Journal of Human Resources 50.2 (2015): 317-372.

Chevalier, Judith, and Glenn Ellison. "Career concerns of mutual fund managers." The Quarterly Journal of Economics114, no. 2 (1999): 389-432.

Greenstone, Michael, and B. Kelsey Jack. "Envirodevonomics: A research agenda for an emerging field." Journal of Economic Literature 53, no. 1 (2015): 5-42.

Greenstone, Michael, John A. List, and Chad Syverson. 2012. "The Effects of Environmental Regulation on the Competitiveness of US Manufacturing." National Bureau of Economic Research Working Paper 18392.

Fan, Haichao, Yu Liu, Larry Qiu, and Xiaoxue Zhao. 2017. "Export to Elude: Evidence from a Tax Enforcement Technology in China." Working Paper, School of Economics Fudan Univeristy.

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

Hanna, Rema. 2010. "US Environmental Regulation and FDI: Evidence from a Panel of US-Based Multinational Firms." American Economic Journal: Applied Economics 2 (3): 158-89.

Hayes, Rachel M., and Scott Schaefer. "Implicit contracts and the explanatory power of top executive compensation for future performance." The RAND Journal of Economics (2000): 273-293.

He, Guojun, and Jeffrey M. Perloff. 2016. "Surface Water Quality and Infant Mortality in China." Economic Development and Cultural Change 65 (1): 119-39.

He, Guojun, and Shaoda Wang. "Do college graduates serving as village officials help rural China?." American Economic Journal: Applied Economics 9, no. 4 (2017): 186-215.

Henderson, J. Vernon. 1996. "Effects of Air Quality Regulation." The American Economic Review 86 (4): 789-813.

Hetts, John J., Michiko Sakuma, and Brett W. Pelham. "Two roads to positive regard: Implicit and explicit self-evaluation and culture." Journal of Experimental Social Psychology 35, no. 6 (1999): 512-559.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." The Quarterly Journal of Economics 124 (4): 1403-48.

Hsieh, Chang-Tai, and Zheng Song. 2015. "Grasp the Large, Let Go of the Small: The

Transformation of the State Sector in China." Brookings Papers on Economic Activity: 328.

Huang, Zhangkai, Lixing Li, Guangrong Ma, and Lixin Colin Xu. Forthcoming. "Hayek, Local Information, and Commanding Heights: Decentralizing State-Owned Enterprises in China." The American Economic Review.

Imbens, Guido W. "Matching methods in practice: Three examples." Journal of Human Resources 50.2 (2015): 373-419.

Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico. "The effects of school spending on educational and economic outcomes: Evidence from school finance reforms." Quarterly Journal of Economics, forthcoming.

Jaffe, Adam B., Steven R. Peterson, Paul R. Portney, and Robert N. Stavins. 1995. "Environmental Regulation and the Competitiveness of US Manufacturing: What does the Evidence Tell Us?" Journal of Economic Literature 33 (1): 132-63.

Ji, Zhihong., Li-an Zhou, Peng Wang, Yingyan Zhao. "Promotion of local officials and bank lending: Evidence from China's city commercial banks." Finance Research Journal (2014): 1-15.

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

Jia, Ruixue, 2017. "Pollution for promotion."

Kahn, Matthew E., and Erin T. Mansur. 2013. "Do local energy prices and regulation affect the geographic concentration of employment?" Journal of Public Economics 101: 105-114.

Kahn, M.E., Li, P. and Zhao, D., 2015. "Water pollution progress at borders: the role of changes in China's political promotion incentives." American Economic Journal: Economic Policy, 7(4), pp.223-42.

Keiser, David A., and Joseph S. Shapiro. 2017. "Consequences of the Clean Water Act and the Demand for Water Quality." National Bureau of Economic Research Working Paper w23070.

Lazear and Oyer The handbook of organizational economics. Gibbons, Robert, and John Roberts, eds. Princeton University Press, 2012.

Lee, David S., and Thomas Lemieuxa. 2010. "Regression Discontinuity Designs in Economics." Journal of Economic Literature 48 (2): 281-355.

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

Lipscomb, Molly, and Ahmed Mushfiq Mobarak. 2017. "Decentralization and pollution spillovers: evidence from the re-drawing of county borders in Brazil." The Review of Economic Studies 84, no. 1 (2016): 464-502.

List, John A., Daniel L. Millimet, Per G. Fredriksson, and W. Warren McHone. 2003. "Effects of Environmental Regulations on Manufacturing Plant Births: Evidence from a Propensity Score Matching Estimator." Review of Economics and Statistics 85 (4): 944-52.

List, J.A. and Sturm, D.M., 2006. "How elections matter: Theory and evidence from environmental policy.” The Quarterly Journal of Economics, 121(4), pp.1249-1281.

Liu, Yu. 2017. “Informal Taxation and Firm Performance: Evidence from China.” Working Paper, School of Economics, Fudan University.

MacLeod, W. Bentley. "Optimal contracting with subjective evaluation." American Economic Review 93, no. 1 (2003): 216-240.

Magruder, Jeremy R. "High unemployment yet few small firms: The role of centralized bargaining in South Africa." American Economic Journal: Applied Economics (2012): 138-166.

Martinez-Bravo, Monica. "The role of local officials in new democracies: Evidence from Indonesia." American Economic Review 104, no. 4 (2014): 1244-87.

Martinez-Vazquez, Jorge, et al. "Expenditure assignments in China: Challenges and policy options." Public Finance in China: Reform and Growth for a Harmonious Society. Shuilin Wang and Jiwei Liu eds. Washington, DC: The World Bank (2008): 77-94.

McCrary, Justin. 2008. “Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test.” Journal of Econometrics 142 (2): 698-714.

Meyer, Margaret, Paul Milgrom, and John Roberts. "Organizational prospects, influence costs, and ownership changes." Journal of Economics & Management Strategy 1, no. 1 (1992): 9-35.

Milgrom, Paul R. "Employment contracts, influence activities, and efficient organization design." Journal of political economy 96, no. 1 (1988): 42-60.

Milgrom, Paul, and John Roberts. "An economic approach to influence activities in organizations." American Journal of sociology 94 (1988): S154-S179.

Munshi, Kaivan, and Mark Rosenzweig. Insiders and outsiders: local ethnic politics and public goods provision. No. w21720. National Bureau of Economic Research, 2015.

Oates, Wallace E. "The effects of property taxes and local public spending on property values: An empirical study of tax capitalization and the Tiebout hypothesis." The Journal of Political Economy (1969): 957-971.

Oates, Wallace E. "Fiscal federalism." Books (1972).

Oates, Wallace E. "Searching for Leviathan: An empirical study." The American Economic Review (1985): 748-757.

Oates, Wallace E. "Searching for Leviathan: A reply and some further reflections." The American Economic Review (1989): 578-583.

Olken, Benjamin A., Rohini Pande, and Raluca Dragusanu. "Governance review paper: J-PAL governance initiative." Rep., Abdul Latif Jameel Poverty Action Lab, Cambridge, MA(2011).

Olley, G. Steven, and Ariel Pakes. 1996. “The Dynamics of Productivity in the Telecommunications Equipment Industry.” Econometrica 64 (6): 1263-1297.

Oyer, P., and S. Schaefer. "Personnel Economics: Hiring and Incentives." Handbook of Labor Economics. (2011): 1769-1823.

Porter, Michael. 1991. “America’s Green Strategy.” Scientific American 264 (4): 168.

Prendergast, Canice. "The provision of incentives in firms." Journal of economic literature 37, no. 1 (1999): 7-63.

Rhode, Paul W., and Koleman S. Strumpf. "Assessing the importance of Tiebout sorting: Local heterogeneity from 1850 to 1990." The American Economic Review 93.5 (2003): 1648-1677.

Richter, Wolfram F., and Dietmar Wellisch. "The provision of local public goods and factors in the presence of firm and household mobility." Journal of Public Economics 60.1 (1996): 73-93.

Rosen, Sherwin. "Hedonic prices and implicit markets: Product differentiation in pure competition." The journal of political economy (1974): 34-55.

Ryan, Stephen P. 2012. "The Costs of Environmental Regulation in a Concentrated Industry." Econometrica 80 (3): 1019-61.

Saavedra, Luz Amparo. "A model of welfare competition with evidence from AFDC." Journal of Urban Economics 47.2 (2000): 248-279.

Sato, Hiroshi. "Public goods provision and rural governance in China." China: An International Journal 6.02 (2008): 281-298.

Schaefer, Scott. "The Dependence of pay—Performance Sensitivity on the Size of the Firm." Review of Economics and Statistics 80, no. 3 (1998): 436-443.

Song, Zheng, Kjetil Storesletten, and Fabrizio Zilibotti. 2011. "Growing Like China." The American Economic Review 101 (1): 196-233.

Shirk, Susan L. The political logic of economic reform in China. Vol. 24. Univ of California Press, 1993.

Suárez Serrato, Juan Carlos, Xiao Yu Wang, and Shuang Zhang. The limits of meritocracy: screening bureaucrats under imperfect verifiability. No. w21963. National Bureau of Economic Research, 2016.

Urquiola, Miguel. "Does school choice lead to sorting? Evidence from Tiebout variation." American Economic Review (2005): 1310-1326.

Van Biesebroeck, Johannes. 2007. "Robustness of Productivity Estimates." The Journal of Industrial Economics 55 (3): 529-69.

Walker, W. Reed. 2011. "Environmental Regulation and Labor Reallocation: Evidence from the Clean Air Act." The American Economic Review 101 (3): 442-47.

Walker, W. Reed. 2013. "The Transitional Costs of Sectoral Reallocation: Evidence from the Clean Air Act and the Workforce." The Quarterly Journal of Economics 128 (4): 1787-835.

Wang, Jin. "The economic impact of special economic zones: Evidence from Chinese municipalities." Journal of Development Economics 101 (2013): 133-147.

White, Michelle J. "Firm suburbanization and urban subcenters." Journal of Urban Economics 3.4 (1976): 323-343.

Wildasin, David E. "Income redistribution in a common labor market." The American Economic Review (1991): 757-774.

Wilson, John Douglas, and Roger H. Gordon. "Expenditure competition." Journal of Public Economic Theory 5.2 (2003): 399-417.

World Bank. 2007. "Cost of Pollution in China: Economic Estimates of Physical Damages." Washington, DC: World Bank.

Xi, Tianyang, Yang Yao, and Muyang Zhang. 2017. "Bureaucratic Capability and Political Opportunism: An Empirical Investigation of City Officials in China." Working Paper, National School of Development, Peking University.

Xing, Chunbing. "Migrate for education? Primary school relocation and migration of rural households." IZA working paper, (2014).

Yang, Rudai. 2015. "Study on the total factor productivity of Chinese manufacturing enterprises." Economic Research Journal, 2, pp.61-74.

Yang, Dongping. "One decade of Rural School Consolidation: An evaluation." Research Report, the 21st Century Education Research Institute, (2012).

Yao, Yang, and Muyang Zhang. "Subnational leaders and economic growth: evidence from Chinese cities." Journal of Economic Growth (2011): 1-32.

Yasar, M., Raciborski, R. and Poi, B. 2008. "Production Function Estimation in Stata using the Olley and Pakes Method." Stata Journal 8(2): 221.

Yu, Miaojie. 2015. "Processing Trade, Tariff Reductions and Firm Productivity: Evidence from Chinese Firms." The Economic Journal 125 (585): 943-88

Zhu, Jieming. "A transitional institution for the emerging land market in urban China." Urban Studies 42.8 (2005): 1369-1390.

Zodrow, George R., and Peter Mieszkowski. "Pigou, Tiebout, property taxation, and the underprovision of local public goods." Journal of Urban Economics 19.3 (1986): 356-370.

Appendix A. Estimation of TFP using Olley-Pakes Method

Our Olley-Pakes TFP measure is constructed based on Brandt et al. (2012) using the Annual Survey of Industrial Firms (ASIF) dataset from 2000 to 2007. We made slight changes to the estimations of some key parameters to improve the accuracy of productivity measurement in the ASIF dataset, as suggested by Yang (2015). We explain in detail how we construct these key parameters.

Gross Output

Following the literature, we use production value, instead of sales, as the gross output measure. Production value and sales differ slightly due to the change in inventories. The former is more closely related to input and productivity, and thus more relevant for TFP estimation.

When constructing output deflators, we follow Yang (2015) by using output price indexes for every 2-digit industry in each year from the “Urban Price Yearbook 2011” published by the National Bureau of Statistics. Because those price indexes are linked across different years, we can use them to deflate yearly nominal output to real output in 2000.

Value Added

When constructing real value added, we subtract from the aforementioned real output the goods purchased for resale, indirect taxes, and material inputs.

We construct input deflators from National Input-Output tables in 1997, 2002, and 2007, to take into account the dynamics of input price in different sectors, as suggested by Yang (2015). By doing so, we are able to deflate nominal inputs in each sector in each year to the real values in 2000.

Employment and Wages

The ASIF dataset contains information on the number of employees and the compensation for labor, including wages, employee supplementary benefits, and insurance. We follow Brandt et al. (2012) to sum up wages, benefits, and insurance as a proxy for total labor compensation.

Capital Stock and Investment

In the ASIF dataset, firms report the value of their fixed capital stock at original purchase prices, as well as capital stock at original purchased prices less accumulated depreciation. Because these values are the sum of nominal values in all the past years, they cannot be taken directly to proxy for real capital stock. To back out the real capital stock and construct real investment from this variable, we follow the approach suggested by Yang (2015).

For each year after the first period, we first take the difference between "current capital stock" and "capital stock in the previous period," then deflate it according to the previously calculated price indexes for this period. For observations in the first period of the panel, we assume that, from the firm's establishment until this first period, it had on average the same increasing trend in investment rate as the 2-digit sector average value, which can be collected from the yearbooks published by the National Bureau of Statistics. Under this assumption, together with the nominal capital stock in the first period, nominal capital stock when established, and relevant deflators, we are able recover the real investment and real capital stock in the first period as well.

TFP Estimation

With the key variables constructed, we follow the literature and use the Olley and Pakes (1996) approach to estimate TFP. This approach addresses both simultaneity and selection problems that are salient in the traditional Solow-residual type TFP estimates. For implementation, we use the Stata package provided by Yasar et al. (2008); please refer to their manual for the details of the estimation.

Appendix B. Conceptual Framework

We provide a conceptual framework that helps to explain the empirical findings. We focus on firms' production decisions and address how environmental regulations can affect their TFP. We assume that firms produce homogeneous goods, with a Hicks-neutral continuously differentiable production function$Q(K, L)$, where K represents capital, L represents labor, and $Q_k, Q_l > 0; Q_{kk}, Q_{ll} < 0$.

When a firm produces output Q, emissions are generated as a by-product and are an increasing function of output Q. The firm can reduce its emissions by employing extra (non-productive) labor L_E and/or capital K_E. The final emission level is therefore a continuously differentiable function $E(Q, K_E, L_E)$. We assume that $E_1 > 0, E_{11} > 0; E_2 > 0, E_{22} < 0; E_3 > 0, E_{33} < 0$ and $E_{23} = E_{32} = 0$.

We model the government's environmental regulations as a unit tax (fine), t, on firm's emissions E. A firm maximizes its profit by setting K, L, K_E, L_E as follows:

(1) $$\max_{K,L,K_E,L_E} \pi = p \cdot Q(K, L) - r \cdot (K + K_E) - w \cdot (L + L_E) - t \cdot E(Q, K_E, L_E)$$

where p represents the market output price, r represents the capital price or interest rate, and w represents wages.

The first order conditions for the firm's profit maximization problem are therefore:

(2) $$\frac{\partial \pi}{\partial K} = p \cdot Q_k - r - t \cdot E_1 \cdot Q_k = 0$$

(3) $$\frac{\partial \pi}{\partial L} = p \cdot Q_l - w - t \cdot E_1 \cdot Q_l = 0$$

(4) $$\frac{\partial \pi}{\partial K_E} = -r - t \cdot E_2 = 0$$

(5) $$\frac{\partial \pi}{\partial L_E} = -w - t \cdot E_3 = 0$$

Applying the implicit function theorem, we can prove the following:

(6) $$\frac{\partial K}{\partial t} < 0, \frac{\partial L}{\partial t} < 0; \frac{\partial K_E}{\partial t} > 0, \frac{\partial L_E}{\partial t} > 0;$$

(7) $$\frac{\partial E / \partial t}{E} < \frac{\partial Q / \partial t}{Q} < 0;$$

(8) $$\frac{\partial Q/\partial t}{Q} < \frac{\partial(K+K_E)/\partial t}{(K+K_E)};\ \ \frac{\partial Q/\partial t}{Q} < \frac{\partial(L+L_E)/\partial t}{(L+L_E)}.$$

Proposition 1. An increase in the emissions tax reduces TFP.

Proof. By definition, $\text{TFP} = \frac{p\cdot Q(K,L)}{r\cdot(K+K_E)+w\cdot(L+L_E)}$; and we therefore have the following:

(11) $$\frac{\partial\text{TFP}}{\partial t} = \frac{\partial[\frac{p\cdot Q(K,L)}{r\cdot(K+K_E)+w\cdot(L+L_E)}]}{\partial t} = p\cdot\frac{r\left[\frac{\partial Q}{\partial t}\cdot(K+K_E)-Q\cdot\frac{\partial(K+K_E)}{\partial t}\right]+w\left[\frac{\partial Q}{\partial t}\cdot(L+L_E)-Q\cdot\frac{\partial(L+L_E)}{\partial t}\right]}{[r\cdot(K+K_E)+w\cdot(L+L_E)]^2} < 0$$

where the inequality follows from Equation (8).

Proposition 2. An increase in the emission tax t reduces the emission level E and emission intensity $\frac{E(Q,K_E,L_E)}{Q}$.

Proof. Taking the derivative of emissions with respect to the emission tax, we have:

(9) $$\frac{\partial E}{\partial t} = E_1\cdot\frac{\partial E}{\partial t} + E_2\cdot\frac{\partial K_E}{\partial t} + E_3\cdot\frac{\partial L_E}{\partial t} < 0;$$

where the inequality follows from Equations (6) and (7).

For emission intensity, we have:

(10) $$\frac{\partial(E/Q)}{\partial t} = \frac{\frac{\partial E}{\partial t}\cdot Q-\frac{\partial E}{\partial t}\cdot E}{E^2} < 0$$

where the inequality follows from Equation (7).

In this model, we implicitly assume that production has no effect on the market price. This assumption is likely to hold in our empirical setting because we focus on a small set of firms concentrated in a small geographical area. On the one hand, these firms face the same market because they are located close to each other; on the other hand, as there are many other firms and buyers in the market, local water quality regulations cannot affect the output market prices. This is important because we cannot directly measure output quantity Q in our firm-level production data. Instead, we can only measure revenue $p\cdot Q(K,L)$. Because firms are price-takers in our setting, we can translate the effects of environmental regulation on revenue-based TFP to real (output-based) TFP. In the case where prices depend on marginal cost, we will underestimate the true TFP effect because the price increases as marginal cost of production increases.

Appendix C. An Order Issued by Kunshan Government to Improve Water Quality around the Monitoring Stations (Scan Copy)

Figure 4.1 Anecdotal Evidence

昆山市两减六治三提升专项行动领导小组办公室

昆 263 办〔2017〕186 号

关于对吴淞江赵屯（石浦）等 3 个断面所属流域工业企业实施全面停产的紧急通知

昆山开发区、昆山高新区、花桥经济开发区管委会，各镇人民政府，市水利局、环保局，水务集团：

近期我市自动监测数据显示，我市国省考断面水质较差，达标形势严峻，尤其是赵屯断面（劣V类）、振东渡口断面（劣 V 类）、千灯浦口断面（V类）均无法达到国省考要求。

为确保我市国省考断面达到国家下达 2018 年度考核目标要求，决定对吴淞江赵屯（石浦）等 3 个断面所属流域工业企业（企业名单附后）自 2017 年 12 月 25 日起至 2018 年 1 月 10 日期间实施全面停产，到期视水质情况，决定是否延期；请相关区镇通知相关企

—1—

措施是否到位；对吴淞江、太仓塘、千灯浦沿线闸门全部关闭，泵站禁止排水；实施停产时、停产期间和复工复产时，请各区镇和相关部门督促企业落实好安全生产各项措施。

停产期间实施日报告制度，市水利局每日对站闸关闭情况、泵站排水情况进行检查，每日下午 4:30 前，将检查结果报市 263 办公室邮箱；各相关区镇每日对相关企业停产、网格员驻厂情况、所辖污水厂排放情况进行检查，每日下午 4:30 前将检查结果报市 263 办公室邮箱。（ks263bgs@163.com）

特此通知，请遵照执行。

附件：吴淞江赵屯（石浦）等 3 个断面所属流域停产企业名单

昆山市"两减六治三提升"专项行动领导小组办公室

2017 年 12 月 24 日

抄报：市委办、市府办。

抄送：市安监局、消防大队。

Kunshan 263 Special Action Team Office

Kunshan 263 Office【2017】 #186

An Urgent Order on Suspending Production of Industrial Enterprises Located in the River Basin of Wusong River Zhaotun (Shipu) Water Quality Monitoring Station and Two Other River Basins

To Kunshan Development Zone, Kunshan High-tech Zone, Huaqiao Economic Development Zone Administrative Board, People's Governments of Townships, Municipal Water Resources Bureau, Municipal Environmental Protection Bureau and Municipal Water Affairs Group:

According to the recent data from the automatic water quality monitoring stations in Kunshan city, the water quality of several river segments used for national and provincial assessment is relatively poor. The situation is particularly severe for the Zhaotun water quality monitoring station (worse than Grade V), Zhengdong ferry station (worse than Grade V), and Qian Deng Pu Kou station (Grade V), all of which may fail to meet the national and provincial assessment requirements.

To ensure that the water quality in these river segments meets the annual national assessment requirement in 2018, we have decided to suspend the production of industrial enterprises (list attached) located near the Wusong River Basin Zhaotun (Shipu) water quality monitoring station and two other river basins near water quality monitoring stations, effective from December 25, 2017 to January 10, 2018. The suspension of production may be further extended, depending on the conditions of water quality readings. Relevant district and township governments should inform the enterprises about the decision. The inspection teams should supervise and take production cessation measures. Special investigators should be placed in the plants to ensure full compliance. Sluice gates along Wusong River, Taicang Embankment and Qian Deng Pu must be closed, and the pumping facilities need to be shut down and stop discharging wastewater during this period. District and township governments and relevant departments shall ensure that enterprises take proper safety measures in the process of suspending and resuming production.

During the production suspension period, a daily reporting system will be adopted. The Municipal Water Resources Bureau shall inspect the status of all sluice gates and pumping facilities and report the inspection results to the City's "263" Office before 4:30pm every day. District and township governments, special investigators based in the plants, and wastewater treatment plants shall check whether there are violations of the production suspension order and report the results to the Office (ks263bgs@163.com) before 4:30 PM every day.

Hereby noticed, and please follow the order.

Appendix: List of industrial enterprises that shall suspend production

The "263" Office of Kunshan Government

24th December, 2017

cc. Municipal Office of Kunshan, Government Office of Kunshan

cc. Kunshan Safety Supervision Bureau, Kunshan City Fire Brigade

Appendix D. Appendix Tables

Table 4.1 Polluting vs Non-Polluting Industries

Polluting Industries		Non-Polluting Industries	
Industry	Code	Industry	Code
Mining and Washing of Coal	6	Forestry	2
Mining and Processing of Ferrous Metal Ores	8	Extraction of Petroleum and Natural Gas	7
Mining and Processing of Non-metallic Mineral	10	Mining and Processing of Non-ferrous Metal Ores	9
Fermentation	14 (6)	Agricultural and Sideline Food Processing	13
Beverage Manufacturing	15	Food Manufacturing	14
Textiles Mills	17	Tobacco Manufacturing	16
Leather, Fur and Related Products Manufacturing	19	Wearing Apparel and Clothing Accessories Manufacturing	18
Pulp and Paper Manufacturing	22 (1, 2)	Wood and Bamboo Products Manufacturing	20
Petrochemicals Manufacturing	25	Furniture Manufacturing	21
Chemical Products Manufacturing	26	Paper Products Manufacturing	22
Medicine Manufacturing	27 (1, 2, 4)	Printing and Reproduction of Recorded Media	23
Chemical Fibers Manufacturing	28	Education and Entertainment Articles Manufacturing	24
Non-Metallic Mineral Products Manufacturing	31	Medical Goods Manufacturing	27
Iron and Steel Smelting	32 (1, 2)	Rubber Products Manufacturing	29
Non-Ferrous Metal Smelting	33 (1)	Plastic Products Manufacturing	30
Fossil-Fuel Power Station	44 (1)	Basic Metal Processing	32
		Non-Ferrous Metal Processing	33
		Fabricated Metal Products Manufacturing	34
		General Purpose Machinery Manufacturing	35
		Special purpose Machinery Manufacturing	36
		Transport Equipment Manufacturing	37
		Electrical Equipment Manufacturing	39
		Computers and Electronic Products Manufacturing	40
		General Instruments and Other Equipment Manufacturing	41
		Craftworks Manufacturing	42
		Renewable Materials Recovery	43
		Electricity and Heat Supply	44
		Gas Production and Supply	45
		Water Production and Supply	46

Notes: Industrial classification for national economic activities (GB/T 4754—2002). The division between polluting Industries and non-polluting Industries is according to the Ministry of Environmental Protection (http://wfs.mep.gov.cn/gywrfz/hbhc/zcfg/201009/t20100914_194483.htm).

Table 4.2 Covariate Balance Between Upstream Townships and Downstream Townships

	Mean		Mean Difference
	Downstream	Upstream	≤10km
	(1)	(2)	(3)
Panel A. Basic Township Characteristics			
Town Area	7,636	7,345	302.02
(Mu)	-4,210	-4,233	(1,467.30)
Arable Area	3,308	2,734	-58.29
(Mu)	-2,005	-1,880	(803.44)
Distance to County Center	2.34	2.45	0.52
(KM)	(0.94)	(1.04)	(0.70)
Old-Region Town	0.21	0.16	-0.14
(1=Old-Region Town)	(0.41)	(0.36)	(0.18)
Minority Town	0.01	0.02	-0.00*
(1=Minority Town)	(0.12)	(0.15)	(0.00)
No. of Residents Communities	1.65	1.29	-2.52
	(4.34)	(2.72)	(4.86)
No. of Villages	25.94	22.03	-1.62
	(16.77)	(14.57)	(5.33)
Panel B. Basic Infrastructure			
Road Length	53.55	44.09	-3.83
(KM)	(46.37)	(43.40)	(11.57)
# of Villages with Paved Road	24.73	21.27	-0.57
	(16.51)	(13.92)	(5.51)
# of Villages with Electricity	25.94	22.03	-1.62
	(16.77)	(14.57)	(5.33)
# of Villages with Tap Water	14.24	10.07	0.69
	(15.65)	(13.60)	(6.02)
Panel C. Human Capital			
No. of Primary School	18.08	17.02	2.10
	(9.38)	(9.45)	(3.99)
No. of Primary School Students	7,271	5,936	-309.76
	-4,824	-3,790	(2,329.78)
Obs.	187	143	104

Notes: Data are collected from the Township Conditions Survey in 2002. Columns 1–2 report the means and standard deviations of township covariates. In columns 3, we test the covariate balance between upstream and downstream towns. The difference coefficients are obtained by running OLS regressions of township variables on an upstream dummy and a set of water quality monitoring station fixed effects. Standard errors reported in the parentheses are clustered at the water monitoring station level.

Table 4.3 Density Tests for Sorting Using Local Polynomial Density Estimation

	(1)	(2)	(3)	(4)
Panel A. Firms 2000-2007, Obs = 6582				
T	0.09	0.73	-12.07	0.09
P>|T|	0.93	0.47	0.00	0.93
Bandwidth Left	2.47	1.97	6.17	2.47
Bandwidth Right	2.01	1.97	6.17	2.01
Panel B. Firms in 2013, Obs = 2562				
T	1.60	1.44	1.26	1.60
P>|T|	0.11	0.15	0.21	0.11
Bandwidth Left	2.93	2.70	3.60	2.93
Bandwidth Right	3.29	2.70	3.60	3.29
Bandwidth Selector	Each	Diff	Sum	Comb

Notes: This table reports RD manipulating tests using the local polynomial density estimators proposed by Cattaneo et al. (2017a, 2007b). We use four different bandwidth selectors to check the robustness of the results. "Each" means we use two distinct bandwidths based on MSE of each density separately for upstream and downstream firms, which is our preferred test. "Diff" bandwidth selection is based on MSE of difference of densities with one common bandwidth. "Sum" bandwidth selection is based on MSE of sum of densities with one common bandwidth. "Sum" selects the largest bandwidth in our tests and creates large noises in the testing results. "Comb" bandwidth is selected as the median of "Each", "Diff" and "Sum". Technical explanations of different bandwidth selectors can be found in Cattaneo et al. (2017a, 2007b).

Table 4.4 Balanced Panel Results

	Polluting Industries			Non-Polluting Industries		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Water Quality Monitoring and TFP						
TFP (log)	0.24	0.28	0.47**	-0.20	-0.22	-0.14
	(0.30)	(0.30)	(0.24)	(0.23)	(0.23)	(0.23)
Bandwidth (km)	4.89	4.55	5.23	3.64	3.29	3.02
Panel B: Water Quality Monitoring and Residual TFP						
TFP (log)	0.38*	0.41**	0.63***	0.05	0.05	0.01
(Station FE Absorbed)	(0.22)	(0.21)	(0.22)	(0.22)	(0.22)	(0.20)
Bandwidth (km)	6.41	6.37	5.63	3.45	3.17	3.42
Panel C: Water Quality Monitoring and Residual TFP						
TFP (log)	0.40*	0.46**	0.32*	0.13	0.12	0.12
(Station and Industry FE Absorbed)	(0.21)	(0.21)	(0.18)	(0.18)	(0.19)	(0.19)
Bandwidth (km)	6.24	5.99	6.12	3.69	3.59	2.99
Obs.	1281	1281	1281	2896	2896	2896
Kernel	Triangle	Epanech.	Uniform	Triangle	Epanech.	Uniform

Notes: This table replicates the baseline results in Table 1.2 with only firms in the balanced panel. Each cell in the table represents a separate regression. TFP is estimated using Olley and Pakes (1996) method. The discontinuities at monitoring stations are estimated using local linear regressions and MSE-optimal bandwidth proposed by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table 4.5 Abatement Facilities

	(1)	(2)	(3)
Panel A: Water Quality Monitoring and Abatement Facilities			
Abatement Facilities	-0.83**	-0.82**	-0.98**
	(0.32)	(0.33)	(0.39)
Bandwidth (km)	3.28	3.29	2.97
Panel B: Residual Abatement Facilities			
Abatement Facilities	-1.09*	-1.03*	-1.26*
(Station FE Absorbed)	(0.60)	(0.58)	(0.69)
Bandwidth (km)	6.09	6.16	4.23
Panel C: Residual Abatement Facilities			
Abatement Facilities	-0.98*	-0.92	-1.17*
(Station and Industry FE Absorbed)	(0.59)	(0.57)	(0.66)
Bandwidth (km)	6.16	6.23	4.40
Obs.	9041	9041	9041
Kernel	Triangle	Epanech.	Uniform

Note: Each cell in the table represents a separate regression. The outcome variable is the amount of abatement facilities owned by the firm. The discontinuities at monitoring stations are estimated using local linear regressions and MSE-optimal bandwidth proposed by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table 4.6 Political Incentive Results with Politician FE

VARIABLES	(1) Log TFP	(2) Log TFP
Downstream	0.35***	0.12
	(0.05)	(0.08)
Distance	-0.01**	-0.01**
	(0.00)	(0.00)
Downstream*Distance	-0.00	-0.01
	(0.01)	(0.01)
Incentive		-0.03
		(0.07)
Downstream*Incentive		0.31**
		(0.11)
Mean of Outcome Variable	2.58	2.58
Station FE	Yes	Yes
Industry FE	Yes	Yes
Leader FE	Yes	Yes
Year FE	Yes	Yes
Observations	19,572	19,572
R-squared	0.361	0.364

Note: This table reports the robustness of the "promotion incentive" results to the inclusion of politician fixed effects. Column 1 presents the baseline coefficients under this specification, column 2 adds the political incentive dummy and its interaction with the downstream dummy. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10% ** significant at 5% *** significant at 1%.

Table 4.7 Polynomial RD Estimates of the Impact of Water Quality Monitoring on TFP

	Polluting Industries					Non-Polluting Industries				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Control for Station FE										
TFP (log)	0.30**	0.33**	0.22	0.38***	0.65***	0.03	0.07	0.02	0.15*	0.19
	(0.12)	(0.16)	(0.21)	(0.12)	(0.19)	(0.07)	(0.10)	(0.15)	(0.09)	(0.14)
R-Square	0.14	0.14	0.14	0.17	0.16	0.08	0.08	0.08	0.09	0.10
Panel B: Control for Station and Industry FE										
TFP (log)	0.22**	0.29*	0.30	0.32***	0.47**	0.00	0.10	0.17	0.13*	0.16
	(0.09)	(0.14)	(0.20)	(0.10)	(0.19)	(0.07)	(0.09)	(0.13)	(0.07)	(0.14)
R-Square	0.26	0.26	0.26	0.28	0.29	0.27	0.27	0.27	0.29	0.27
Obs.	6,582	6,582	6,582	4,462	1,474	12,422	12,422	12,422	8,981	3,260
Polynomial Function	Linear	Quadratic	Cubic	Linear	Linear	Linear	Quadratic	Cubic	Linear	Linear
Sample	20km	20km	20km	10km	5km	20km	20km	20km	10km	5km

Notes: Each cell in the table represents a separate regression. TFP is estimated using Olley and Pakes (1996) method. We report OLS estimates of the coefficient on a "downstream" dummy after controlling for polynomial functions in distance from the water quality monitoring stations interacted with a downstream dummy. Standard errors clustered at the monitoring station level are reported below the coefficients. * significant at 10% ** significant at 5% *** significant at 1%.

Table 4.8 RD Estimates of the Impact of Water Quality Monitoring on Alternative TFP

	Polluting Industries			Non-Polluting Industries		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Water Quality Monitoring and TFP						
TFP (log) - Polluting Industries	0.61*	0.67*	0.83**	-0.06	-0.10	-0.05
	(0.35)	(0.37)	(0.39)	(0.21)	(0.22)	(0.25)
Panel B: Water Quality Monitoring and Residual TFP						
TFP (log) - Polluting Industries	0.55**	0.54**	0.69**	0.07	0.06	0.00
(Station FE Absorbed)	(0.27)	(0.27)	(0.31)	(0.13)	(0.13)	(0.14)
Panel C: Water Quality Monitoring and Residual TFP						
TFP (log) - Polluting Industries	0.31	0.35	0.57*	0.15	0.14	0.08
(Station and Industry FE Absorbed)	(0.23)	(0.24)	(0.30)	(0.12)	(0.13)	(0.12)
Obs.	6,039	6,039	6,039	11,440	11,440	11,440
Kernel	Triangle	Epanech.	Uniform	Triangle	Epanech.	Uniform

Notes: Each cell in the table represents a separate regression. TFP is estimated using method proposed by Ackerberg et al. (2015). The discontinuities at monitoring stations are estimated using local linear regression and MSE-optimal bandwidth proposed by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods. * significant at 10% ** significant at 5% *** significant at 1%.

Table 4.9 Effect of Monitoring on TFP and Emission by Size

	Small Firms/Emitters			Large Firms/Emitters		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: RD Estimates for TFP (log)						
Downstream	0.11	0.11	0.24	0.31**	0.32**	0.30
	(0.17)	(0.18)	(0.17)	(0.14)	(0.16)	(0.19)
Obs.	3,038	3,038	3,038	3,538	3,538	3,538
Panel B: RD estimates for COD Emission Intensity (log)						
Downstream	0.49	0.44	0.25	0.92*	0.80*	0.48
	(0.40)	(0.36)	(0.28)	(0.47)	(0.45)	(0.44)
Obs.	4,901	4,901	4,901	4,906	4,906	4,906
Kernel	Triangle	Epanech.	Uniform	Triangle	Epanech.	Uniform

Notes: Each cell in the table represents a separate regression. Monitoring station and industry fixed effects are absorbed before estimating regression discontinuity. In columns 1–3, we report the estimated discontinuity for smaller firms or emitters, and in columns 4–6, we report the estimated discontinuity for large firms or emitters. Local linear regression and MSE-optimal bandwidth proposed by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods are used for the estimation. * significant at 10% ** significant at 5% *** significant at 1%.

Table 4.10 Effect of Water Quality Monitoring on Firm Exit

	Exit – Polluting Industries			Exit – Non-Polluting Industries		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Conventional RD Estimates						
Downstream	0.07	0.07	0.11	-0.02	-0.02	-0.04
	(0.09)	(0.09)	(0.10)	(0.05)	(0.05)	(0.05)
Panel B: Bias-Corrected Robust RD Estimates						
Downstream	0.10	0.10	0.15	-0.01	-0.02	-0.04
	(0.10)	(0.10)	(0.11)	(0.07)	(0.07)	(0.06)
Obs.	6,581	6,581	6,581	12,422	12,422	12,422
Kernel	Triangle	Epanech.	Uniform	Triangle	Epanech.	Uniform

Notes: Each cell in the table represents a separate regression. Monitoring station and industry fixed effects are absorbed before estimating regression discontinuity. In columns 1–3, we report the estimated discontinuity for polluting industries, and in columns 4–6, we report the estimated discontinuity for non-polluting industries. Local linear regression and MSE-optimal bandwidth proposed by Calonico et al. (2014) and Calonico (2017) for different kernel weighting methods are used for the estimation. Conventional local linear regression discontinuity standard errors clustered at the monitoring station level are reported below the estimates in Panel A; bias-corrected RD estimates and robust standard errors are reported in Panel B. * significant at 10% ** significant at 5% *** significant at 1%.

APPENDIX REFERENCES

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer. 2015. Identification Properties of Recent Production Function Estimators. Econometrica, 83(6), pp.2411-2451.

Brandt, L., Van Biesebroeck, J. and Zhang, Y. 2012. "Creative Accounting or Creative Destruction? Firm-Level Productivity Growth in Chinese Manufacturing." Journal of Development Economics, 97(2), pp.339-351.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. Forthcoming. "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference." Journal of the American Statistical Association.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." Econometrica 82 (6): 2295-326.

Olley, G. Steven, and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." Econometrica 64 (6): 1263-1297.

Yang, Rudai. 2015. "Study on the Total Factor Productivity of Chinese Manufacturing Enterprises." Economic Research Journal (Chinese), 2, p.006.