THREE ESSAYS ON STRATEGIC MISREPORTING

by

Chun Song

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Agricultural Economics

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Juan Sesmero, Chair
Department of Agricultural Economics

Dr. Michael S. Delgado
Department of Agricultural Economics

Dr. Tim Cason
Department of Economics

Dr. Steven Yu-Ping Wu
Department of Agricultural Economics

Approved by:
Dr. Nicole J. O. Widmar

To my parents, my cat, my professors, and my friends,

I could not have completed this work without you (the order does not necessarily indicate the magnitude of contribution)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This dissertation studies the economics of strategic misreporting and the effect of different anti-misreporting approaches based on theoretical, experimental, and quasi-experimental evidence. In Essay 1, I propose a theoretical model to study the efficacy of absolute and relative inspection standards in reducing misreporting when agents are heterogeneous in their reporting cost. I extend from previous theoretical studies by examining explicitly the performance of competitive endogenous audit rule (i.e., tournament audit) compared to the random audit as a function of agent's heterogeneity parameter. I find that a tournament audit reduces average misreporting and the dispersion of misreporting relative to a random audit, and that the magnitude of the reduction is independent of the degree of heterogeneity among agents. A larger number of audits (presumably delivered by a softer budget constraint), a higher degree of imperfect monitoring, and larger risk aversion among agents reduce the effectiveness of the tournament audit in lowering misreporting. However, the magnitude of the reduction remains independent of heterogeneity in those cases.

Theoretical predictions from the first essay are built on a strategic equilibrium concept that relies on rather sophisticated assumptions. Testing these predictions in a controlled environment is thus of empirical importance. In Essay 2, I study misreporting decisions in laboratory experiments, and I test predictions from the first essay. The game played by subjects carefully recreates the environment used to generate theoretical predictions. The experiments have two sources of exogenous variation. The first varies the audit scheme, while the second varies heterogeneity in the cost of reporting. This allows me to test the key predictions from Essay 1 by comparing outcomes across different combinations of treatments. The experimental results largely support the theoretical predictions that a tournament audit reduces misreporting, both with homogeneous and heterogeneous agents. It also supports the prediction that the magnitude of the reduction in misreporting under a tournament audit relative to the random audit is largely independent of the degree of heterogeneity. However, the misreporting reduction is smaller than predicted, as subjects in the experiment tend to misreport less in the random audit baseline. This result is consistent with subjects being risk averse as characterized in Essay 1. Similarly, efficiency gains associated with lower misreporting are smaller than predicted.

In the third essay, I study a reform that conferred Chinese provincial authorities more monitoring power over air pollution performance by cities in those provinces. I use quasi-experimental methods to quantify the effects of this reform on misreporting by local authorities. Implemented in 2016, the reform gave the provincial authorities direct access to local (municipal) pollution monitoring stations, thereby making it harder for local authorities to misreport after the reform. The reform was introduced only in some provinces, many treated and untreated provinces have similar pollution trends before the reform and significant overlap on observable characteristics. These features aid me in establishing a causal effect of the reform on misreporting. The estimation involves two steps. First, I quantify different types of misreporting following recently proposed methodologies. Second, I regress estimated misreporting on the reform indicator using a difference in difference estimator. I found that the reform reduces hard misreporting, which takes place when local authorities interfere with the pollution monitoring facility, both during regular days and during heavily polluted days. The reform does not appear to reduce soft misreporting, which takes place when local authorities tamper with the pollution data. The results are robust to a number of robustness tests, and suggest that through proper institutional reform, the upper-level government can prevent certain types of misreporting at the local level.

This dissertation delivers a characterization of strategic misreporting by heterogeneous agents and studies the impact of different anti-misreporting schemes based on theoretical, experimental, and observational evidence. Results from this dissertation provide evidence that regulators can use mechanisms that: 1) curb misreporting without enhancing monitoring (tournament audits), or 2) that enhance monitoring to ultimately curb misreporting (adoption of monitoring technologies), or 3) a combination of both. This is important given the pervasiveness of misreporting among regulated agents, and substantial heterogeneity among those agents.

ESSAY 1. STRATEGIC MISREPORTING UNDER ALTERNATIVE AUDIT MECHANISMS WITH HETEROGENEOUS AGENTS

Abstract

In this paper I study the effect of alternative auditing schemes on misreporting by heterogeneous agents. Agents produce an output (pollution) and must report the level of that output, which is not directly observable by the principal. Agents incur additional costs when they report higher pollution, so they have incentives to underreport. However, if audited, they are punished for underreporting pollution. In contrast to a random audit, a tournament audit creates a situation whereby the probability of being audited raises with the magnitude of misreporting, all else constant. I consider agents that vary on their reporting cost: some agents have high reporting cost and some low reporting costs. I find that, when heterogeneity is mean preserving, a tournament audit reduces misreporting relative to a random audit, and that the magnitude of the reduction in *average* misreporting is independent of the degree of heterogeneity among agents. A tournament audit also reduces *dispersion* in misreporting. It does so by reducing misreporting of the high-cost agent more than the low-cost agent. This is a desirable property of the tournament audit because, in an environmental regulation setting, agents that have a higher cost of misreporting also tend to be those whose pollution causes more damage.

1 Introduction

Economic actors are often required to report privately observed information. Such reported information is usually linked to a payoff. Examples of this abound, but a prominent one (one that constitutes a common theme of this dissertation) is when local government agencies (agents) must report economic or environmental metrics to a higher authority (principal), and important fiscal and administrative decisions are tied to those metrics. Tying rewards/penalties to reported performance creates incentives for bureaucracies to misreport and overstate performance, however measured. Moreover, those incentives need not be uniform across agents; agents typically differ in their cost of reporting or benefits from misreporting. Principals implement audits to detect misreporting but are limited by budget constraints and can only audit a small fraction of agents. It is, therefore, crucially important to identify audit schemes that are effective in reducing

misreporting by heterogeneous agents, under scarce audit resources. In this paper, I study the relative effectiveness of two alternative audit schemes–random audit and tournament audit–when the principal interacts with heterogeneous agents.

Previous studies compared misreporting under random and tournament audits and found that a tournament audit reduces misreporting (Gilpatric et al., 2011; Gilpatric et al., 2015; Cason et al., 2016). These studies have investigated this question in a setting with homogeneous agents. However, the reporting costs for different agents are often inherently heterogeneous. For example, in the public sector, some agents (e.g., large cities; cities that do not align politically with the principal) face more serious consequences than others for disclosing truthfully, thus bearing higher reporting costs. In the private sector, polluting firms with little political power may also bear higher reporting costs. Similarly, large firms may be better able to shoulder the financial burden associated with reporting costs, thereby facing lower reporting costs. Sometimes differences in reporting costs are built into regulatory rules. For instance, in 2003 the Environmental Protection Agency (EPA) updated Clean Water Act regulations to subject larger livestock operations (called Concentrated Animal Feeding Operations, CAFOs) to more stringent pollution control than their smaller counterparts (called Animal Feeding Operations, AFOs). In other situations, the EPA sets different per unit emission costs for different agents (EPA, 2020).

Intuitively, the introduction of heterogeneity has both direct and indirect effects on agents' misreporting decisions. First, by construction, heterogeneity raises the reporting cost of some agents and reduces the reporting cost of others. A key issue is whether this affects average and, thereby, aggregate misreporting. This effect takes place both in random and tournament audit schemes. A second, indirect effect is due to the competition created by the tournament scheme. The tournament scheme incentivizes agents to misreport strategically against other agents, striking a balance between the cost of truthful reporting and the expected penalty from misreporting. In this study, I aim to characterize these forces, their interaction, and how they affect the effectiveness of tournament relative to random auditing schemes.

I develop a game in which multiple agents decide on output and then reporting on the output, whereas the principal implements an audit scheme to discourage output misreporting. Following previous studies (Cason et al., 2016) we use a framework in which the audit scheme does not affect output itself, but it affects reporting behavior. The agent chooses the output and observes it privately, they pay the cost for the output voluntarily reported, and they face a chance

that they will be inspected by the supervisor (the principal), in which case any misreporting is detected and punished. The principal implements either a random audit or a deterministic, winner-takes-all type of audit that has a similar structure to the rank-order tournament (Lazear & Rosen, 1981). In the tournament audit, agents act strategically. In other words, each agent knows her own as well as her opponent's reporting costs but does not know others' misreporting at the time of their reporting decision. Crucially for this study, I allow heterogeneity in agents' reporting cost. I assume heterogeneity is mean preserving, i.e., while an increase in heterogeneity raises the spread in reporting cost across agents, the average reporting cost remains constant. When heterogeneity is not mean preserving, the effect of heterogeneity on misreporting is mechanical, not behavioral. By implementing mean preserving increases in heterogeneity, we can focus on the behavioral implications of heterogeneity.

I find that the tournament audit halves average misreporting and the dispersion of misreporting. The magnitude of the reduction is independent of the degree of heterogeneity among agents. This indicates that previous insights from studies with homogeneous agents should, theoretically, extend to a setting with heterogeneous agents. The reduction in average misreporting is higher for the underdog (the agent with high reporting cost) than for the favorite (the agent with the low reporting cost). Importantly, the dispersion of misreporting is also smaller with the tournament audit, suggesting that the tournament scheme can be used to discourage extreme misreporting of the underdog, which is likely to be a key concern of the regulator in practice. The mechanism behind the better performance of tournament audit is the use of the pollution estimation. In the tournament audit, the supervisor uses the noisy proxy of the true pollution to decide which agents to audit. Although imprecise, with multiple agents this estimation is unbiased on average and help improve better selection of dubious reported pollution. In addition, I also study the agents' net payoff under different audit schemes. The numerical results show that the net payoff is about 7% higher for all the agents under the tournament audit scheme regardless of the agents' heterogeneity. This indicates that the tournament audit not only improves efficiency (lower misreporting) but also improves net payoff for the agents.

A larger number of audits (presumably delivered by a softer budget constraint) reduces the effectiveness of the tournament audit in lowering misreporting relative to the random audit. To incentivize the same low level of equilibrium misreporting as in the tournament audit, the random scheme must increase the number of audits by 45%. In other words, the audit rate in the random

audit has to increase from 50% (the assumed baseline audit rate) to 73%. A higher degree of imperfect monitoring (less precise estimation of the pollution), and larger risk aversion among agents also reduce the effectiveness of the tournament audit in lowering misreporting relative to the random audit. However, the magnitude of the reduction remains independent of heterogeneity.

This paper contributes to the literature on relative performance evaluation. There has been a long-lasting debate over absolute and relative performance evaluation standards. With absolute standards, the incentives are assigned in a fixed manner. With relative standards, the incentives are assigned by comparing individual performance against the performance of peers. Previous studies have examined how heterogeneity affects the performance of tournaments relative to the homogeneity case, but the comparison is between piece rate and tournament. Since a random audit is fundamentally different from piece rate, the insights from those studies are not directly applicable to the case of random versus tournament audit (Tsoulouhas & Marinakis, 2007; Ryvkin, 2009; Schotter & Wright, 1992; O'Keeffe et al, 1984; Baik, 1994). Other studies have compared random and tournament audits, but in a setting where agents are homogeneous. To the best of my knowledge, this is the first article that compares the effects of different audit schemes with heterogeneous agents using a formal, yet tractable theoretical model.

Introducing heterogeneity allows me to study the effect of heterogeneity itself on average and aggregate misreporting, as well as the dispersion of misreporting among different agents. But it also allows me to examine how heterogeneity interacts with the audit scheme. I do so by studying a "difference-in-difference" model of heterogeneity. Specifically, I vary the level of heterogeneity and assess the relative effectiveness of the tournament at this new level of heterogeneity, i.e., when agents are heterogeneous, what is the difference between tournament and random audit relative to a setting with homogeneous agents. Previous studies have examined this difference-in-difference but, again, not in the context of audits, but of payment schemes. Since a piece rate scheme is not equivalent to a random audit, insights from those papers do not apply to audit schemes.

This research extends our understanding of competing audit schemes to settings with heterogeneous agents. Although much of the analysis is motivated by lower-level bureaucracies reporting on pollution output to higher-level bureaucracies, my insights also apply to broader regulatory circumstances where heterogeneity is widely present and could play an important role in affecting the agent's decisions. These circumstances include enforcement of environmental

regulations and disclosure of firm emissions, banking regulations, individual tax compliance, among others.

2 Literature Review

The literature on the use of tournaments as incentive schemes consists of two relatively fragmented strands. One strand focuses on the use of tournaments to enhance performance, which we will henceforth call standard tournament literature. Another strand focuses on the use of tournaments to deter misreporting of performance, which we will henceforth call tournament audit literature. In Figure 1-1, we offer a synthesis of this literature.

A standard tournament is a situation in which agents' reward (or penalty) depends on their performance relative to their competitors; and their performance is, at least in part, influenced by their effort. This structure is designed to incentivize effort among agents looking to maximize their expected rewards or minimize their expected penalties. Standard tournaments vary in two key dimensions: how effort translates into performance (production function), and how performance affects the probability of winning (contest success function). These variations generate all-pay auctions, tournaments (Lazear & Rosen, 1981), and lottery contests (Tullock, 1980), among others.

Within the standard tournament literature, some studies have compared outcomes when rewards are determined based on relative performance (tournaments) with outcomes when rewards are determined based on absolute performance (for example, a piece-rate payment scheme). But these studies do not explore the impact of heterogeneity among agents. These include Lazear and Rosen (1981), Holmstrom (1982), and Green and Stockey (1983). Other studies have instead focused on the effect of heterogeneity on performance in a tournament (Schotter and others including Ryvkin). But these studies did not compare schemes where rewards are based on relative and absolute performance.

In contrast to other studies in the standard tournament literature, Tsoulouhas and Marinakis (2007) did compare relative and absolute performance schemes while considering heterogeneous agents. However, the standard tournament literature differs from the audit tournament literature in two crucial ways. First, the baseline in a standard tournament literature–a piece-rate payment scheme–is fundamentally different from the baseline in the audit tournament literature–random audit. Second, in an audit tournament the level of misreporting not only affects who gets audited, but the penalty is conditional upon being audited. This differs from the standard tournament in

which the effort only affects the chance of winning but not the amount of the prize. Because of these differences, insights from Tsoulouhas and Marinakis (2007) are not generalizable to the audit tournament context.

The studies in the tournament audit literature have compared schemes with rewards based on relative misreporting (tournaments) vis-à-vis schemes with rewards based on absolute misreporting (random auditing scheme). Theory and experiments in this literature have found evidence that tournament audit schemes reach higher levels of disclosure relative to random audits. The intuition is that, in addition to the typical incentives provided by a fixed audit probability and penalties conditional on being caught, a tournament audit provides a further incentive to report truthfully, as this decreases the audit probability. However, these studies have not considered heterogeneous agents, for example Gilpatric et al (2011), Gilpatric et al (2015), and Cason et al (2016). Therefore, three questions remain unanswered: 1) how does heterogeneity affect reporting conditional on the audit scheme? 2) how does a tournament audit scheme affect misreporting conditional on heterogeneity? and 3) how do heterogeneity and the tournament audit scheme interact to shape misreporting?

A key issue in our study is how to specify heterogeneity. This is important because, if heterogeneity affects average cost of reporting, then it would mechanically impact misreporting. In fact, much of the literature specifies heterogeneity in a way that is not mean preserving. For example, Schotter & Wright (1992) and Ryvkin (2009) introduce a type of heterogeneity that raises average cost of shirking and, unsurprisingly, find that heterogeneity enhances performance under a piece-rate payment scheme. And since heterogeneity increases baseline performance, a tournament scheme has a smaller effect on performance.

Tsoulouhas & Marinakis (2007) consider mean-preserving heterogeneity while comparing tournament and piece-rate schemes. They measure heterogeneity by the variance of the ability distribution. With larger variance, agents are more heterogeneous, and relative performance evaluation via tournaments is less desirable than absolute performance evaluation via piece rate. Moreover, they also find that the larger the variances in idiosyncratic shock, the lower the advantage of the piece rate because of the weaker link between the power of incentives and output. Similarly, Moldovanu & Sela (2006) investigated heterogeneity from the contest designer's point of view. The structure is an all-pay auction where contestants have abilities drawn from a distribution. They show that splitting the competitors into two divisions is optimal if the designer's

goal is to maximize the highest rather than the aggregate effort. Also, Ryvkin (2009) shows that even a small deviation from symmetry (in the sense of agent's ability) results in inefficiency of the tournament. According to Kräkel and Schöttner (2010), the motivating effect of relative performance evaluation with heterogeneous workers depends on whether workers' abilities and the firm's production technology are complementary. O'Keeffe et al (1984) first raised the question of how agents might behave in an uneven tournament in which one of the agents is "disadvantaged" in terms of ability. Later, Schotter & Weigelt (1992) tested the case where the rules favor one player over another and in the case where players have different costs. They find that both asymmetries reduce individual efforts. Alm & McKee (2004) investigated coordination in a misreporting game. Harbing et al (2007) studied the impact of asymmetry in a contest with the possibility of sabotage and found that different equilibria can appear under different compositions of the player's type. Heterogeneity is also found to increase the intensity of sabotage against abler players in contests (Chen, 2003).

In sum, agents' heterogeneity has been studied extensively in standard contests, but not in tournament audits. Incorporating heterogeneity in tournament audits is key in light of substantial diversity in agents' characteristics both in public and private settings. I develop a model that is designed to answer the three unanswered questions previously discussed. We now turn to the model and the nature of game introduced by a tournament audit with heterogeneous agents.

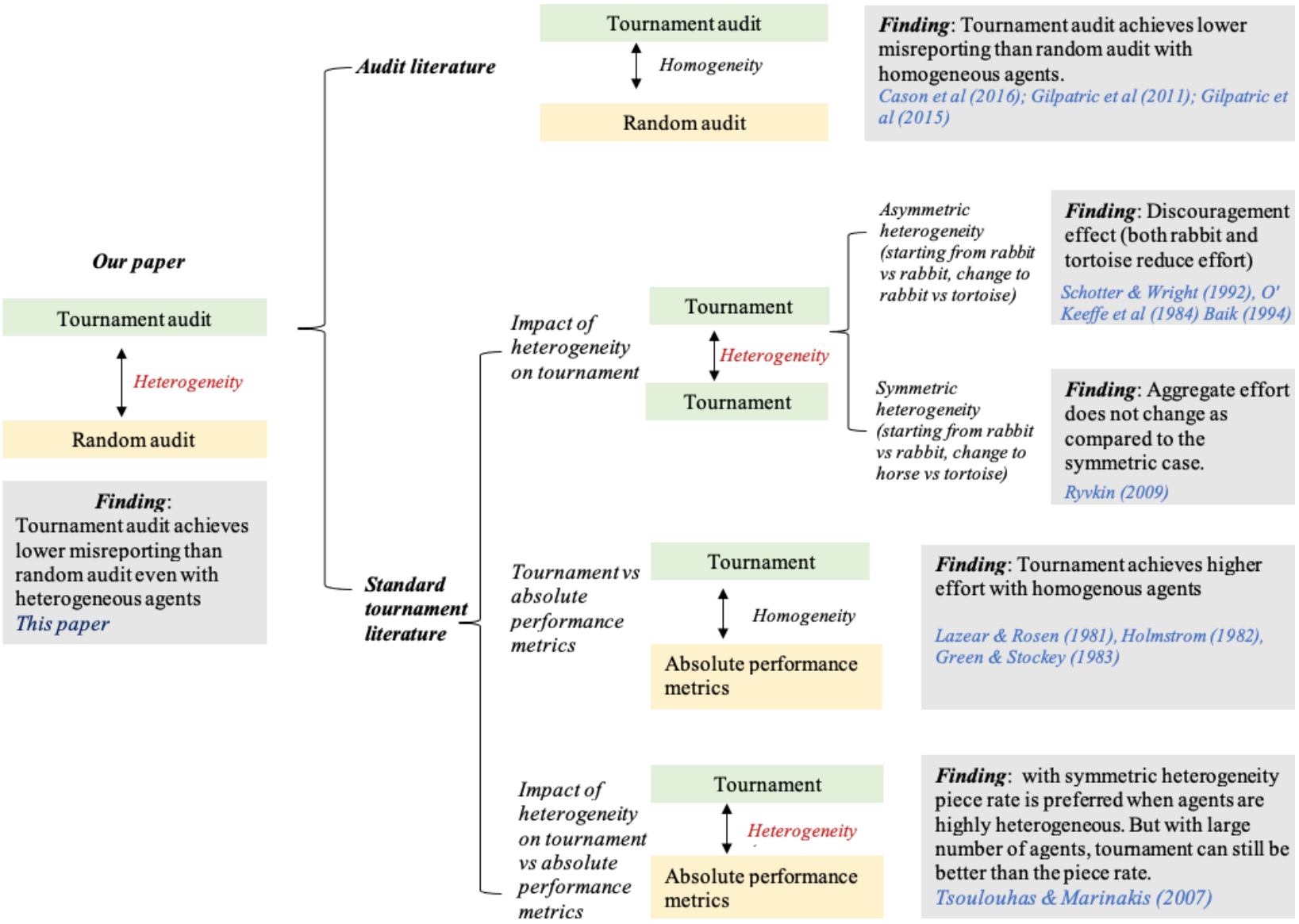


Figure 1-1 Literature summary

3 Theoretical Model

In this section, I introduce the theoretical model. The game is organized as a multi-agent, static, simultaneous-move game with complete information. Consider four risk-neutral agents indexed by $i = 1, 2, 3, 4$.[1] Each agent individually produces an output y_i and reports r_i to the supervising entity (the principal). The reported output may not necessarily be equal to the actual output and the gap between output and reporting is defined as misreporting, or $\Delta_i = \mathrm{y} - r_i$. As an example, the output can be thought of as pollution, and the reporting can be thought of as the reported level of pollution.

The agent receives a private benefit $B(y_i)$ that is increasing and concave in its output and incurs a production cost $C_1(y_i)$ that is increasing and convex in output. When reporting the level of output r_i to their supervisor, the agent incurs a per-unit reporting cost t. In the context of government's environmental performance, the cost can be thought of as the penalty for environmental degradation. In the context of firm-level emissions, it could be a pollution tax.

There are two types of agents, denoted as favorites and underdogs. I assume there are four agents in the model: two favorites and two underdogs. This assumption delivers an even composition of types while, at the same time, ensuring that each type interacts with more than one other agent. In section 6.2 I discuss the case where the group displays an uneven mix of player types. The favorite pays a lower cost than the underdog per unit of reported output.

The way we specify heterogeneity is motivated by many real-world examples. Here we provide two examples in the context of pollution emission and environmental quality monitoring. The first example is related to how different cities in China might have different cost to disclose the true pollution level. In China, cities with better environmental quality are recognized and rewarded with the title *Environmental Quality Demonstration City*. These are honorable titles representing that these cities are environment friendly. They are also regarded as the role models for the rest of the cities. For these environment friendly cities, to disclose the same level of pollution induces a much higher reputational cost. In this example, these cities are similar to the underdog in our model because it is more expensive for these cities to report pollution truthfully. The second example related to our model is how the location of farm can affect the cost to report fertility application. For example, for the same amount of applied fertility, the farm that is closer

[1] In section 6.3 I relax this assumption and discuss the case with four risk averse agents.

to the river has higher leakage and may face higher cost due to the potential pollution charge. For these farms, the cost of reporting truthfully the fertility used is higher and are similar to the underdog in our model.

The supervisor does not observe y_i but can conduct the inspection (audit), which reveals the true output with certainty.[2] Because of budget constraints, I assume that the supervisor audits two out of the four agents. In section 6.2 I examine the implications of relaxing this assumption. The probability that an individual agent is audited, $Prob(agent\ i\ is\ audited)$, depends on the audit rules. I consider two schemes: a random audit scheme and a tournament audit scheme.

In the random audit, the supervisor audits two agents randomly so the chance for each agent to be audited is

$$Prob(agent\ i\ is\ audited) = 1/2$$

In the tournament audit, the supervisor has a noisy estimation of the output $\hat{y}_i = y_i + \varepsilon_i$, where ε_i is a random variable representing noise in estimation. I assume the noise ε has mean zero. This is a plausible assumption because the supervisor might over-estimate or underestimate the output of a single agent, but due to knowledge or experience and imperfect proxies it is likely to have an unbiased estimation of output. It is also assumed that the audit noise ε is uniformly *i.i.d* over the support of $[-q, q]$. In effect, the distribution of ε_i captures the degree of imperfect monitoring, the larger the support, the less precise the estimation.

If the supervisor detects any misreporting by agent i, it imposes a fine $F(y_i - r_i)$ on the agent, which includes the output cost unpaid due to misreporting plus a penalty that is convex and increasing in the magnitude of misreporting. In practice the penalty represents the political, judicial, or financial penalty for detected manipulation.[3]

The supervisor ranks the agents based on the gap between each agent's reporting and the noisy output observed by the supervisor. The supervisor audits the top two agents with the largest gap or $r_i - \hat{y}_i = r_i - (y_i + \varepsilon_i) = \Delta_i - \varepsilon_i$. Therefore, agent i's probability of being audited is a function of all agents' output and misreporting decisions. I denote the probability that the favorite is audited as $P^f(r_f, y_f, r_u, y_u)$ and the probability that the underdog is audited as $P^u(r_f, y_f, r_u, y_u)$. I

[2] In the standard contest, principal may be allowed to set the contest parameters such as the prize spread. Here the principal's behavior is entirely modeled through the tournament audit Abstracting from the case when principal chooses monitoring parameters allows us to examine the full impact of heterogeneity on agent's decision.

[3] For example, the Valencian district in Spain was found by European Council to had underreported the debt to GDP rate in 2012 and is imposed a fine of € 18 million (European Council, 2015). In China, several local government officials who were detected to have falsified pollution readings were removed from the position (Ma Y., 2017).

derive formal expressions for these probabilities in Appendix A.1. All else constant, each agent's probability of being audit is increasing in the degree of underreporting, i.e., increasing in output conditional on reporting and decreasing in reporting conditional on output. Formally, this implies that $\frac{\partial P^f(r_f,y_f,r_u,y_u)}{\partial r_f} = -\frac{\partial P^f(r_f,y_f,r_u,y_u)}{\partial y_f}$ and $\frac{\partial P^u(r_f,y_f,r_u,y_u)}{\partial r_u} = -\frac{\partial P^u(r_f,y_f,r_u,y_u)}{\partial y_u}$.

Because the audit agency faces a budget constraint and can not audit all agents, everyone can still misreport if they simply make sure that their susceptibility is lower than the other's. Intuitively, under this mechanism, the agent's misreporting raises the probability of audit (in contrast to the random auditing scheme), all else constant. This should weaken the incentives for misreporting. However, the strength of this force depends upon the agent's beliefs regarding other agents' misreporting, which is influenced by the other agents' characteristics. This is a key mechanism in the case of heterogeneity.

Having defined all the major components of the model, I now characterize the objective functions for each agent under different audit rules. With the random audit, I denote it using the superscript Rdm. The objective functions for the favorite (with the subscript f) and the underdog (with the subscript u), respectively, are:

$$z_f^{Rdm} = B\left(y_f^{Rdm}\right) - C_1\left(y_f^{Rdm}\right) - (t-\gamma)r_f^{Rdm} - 0.5F\left(y_f^{Rdm} - r_f^{Rdm}\right) \quad (1)$$

$$z_u^{Rdm} = B(y_u^{Rdm}) - C_1(y_u^{Rdm}) - (t+\gamma)r_u^{Rdm} - 0.5F(y_u^{Rdm} - r_u^{Rdm}) \quad (2)$$

Where 0.5 represents the random audit probability. The first order conditions for the favorite with respect to the output and reporting decisions are shown in equation (3) and (4). The solution is the same whether the system is solved simultaneously (with agent chooses output and report simultaneously) or sequentially (with agent first chooses output and then report). This is because the expected penalty depends only on the magnitude of the gap between the output and report, but not on report itself. Thus, when choosing the report, the agent is choosing the amount of misreporting.

$$\frac{\partial z_f^{Rdm}\left(y_f^{Rdm}, r_f^{Rdm}, y_u^{Rdm}, r_u^{Rdm}\right)}{\partial y_f^{Rdm}} = B_1'\left(y_f^{Rdm}\right) - C_1'\left(y_f^{Rdm}\right) - 0.5F'\left(y_f^{Rdm} - r_f^{Rdm}\right) = 0 \quad (3)$$

$$\frac{\partial z_f^{Rdm}\left(y_f^{Rdm}, r_f^{Rdm}, y_u^{Rdm}, r_u^{Rdm}\right)}{\partial r_f^{Rdm}} = -(\mathrm{t}-\gamma) + 0.5F'\left(y_f^{Rdm} - r_f^{Rdm}\right) = 0 \quad (4)$$

The FOC for the underdog are:

$$\frac{\partial z_u^{Rdm}\left(y_f^{Rdm}, r_f^{Rdm}, y_u^{Rdom}, r_u^{Rdm}\right)}{\partial y_u^{Rdm}} = B_1'(y_u^{Rdm}) - C_1'(y_u^{Rdm}) - 0.5F'(y_u^{Rdm} - r_u^{Rdm}) = 0 \quad (5)$$

$$\frac{\partial z_u^{Rdm}\left(y_f^{Rdm}, r_f^{Rdm}, y_u^{Rdom}, r_u^{Rdm}\right)}{\partial r_u^{Rdm}} = -(\mathrm{t} + \gamma) + 0.5F'(y_u^{Rdm} - r_u^{Rdm}) = 0 \quad (6)$$

Notice that these first order conditions of the favorite and the underdog are independent. Agents do not act strategically under a random audit rule, so the first order conditions do not characterize the best response functions and they are solved independently.

Since $0.5F'\left(y_f - r_f\right)$ are shown in both the FOC with respect to the output and the reporting variable for the favorite and the same is true for the underdog, it is straightforward to show that the solution of the output satisfies:

$$B_1'\left(y_f^{Rdm}\right) - C_1'\left(y_f^{Rdm}\right) = t - \gamma \quad (7)$$

$$B_1'\left(y_u^{Rdm}\right) - C_1'\left(y_u^{Rdm}\right) = t + \gamma \quad (8)$$

I denote control variables in the tournament audit by the superscript Tnmt. l also assume that the agents choose their reporting levels simultaneously and employ a Nash equilibrium as the solution concept. The objective functions are:

$$z_f^{Tnmt} = B\left(y_f^{Tnmt}\right) - C_1\left(y_f^{Tnmt}\right) - (t - \gamma) * r_f^{Tnmt} - P^f\left(r_f^{Tnmt}, y_f^{Tnmt}, r_u^{Tnmt}, y_u^{Tnmt}\right)F\left(y_f^{Tnmt} - r_f^{Tnmt}\right) \quad (9)$$

$$z_u^{Tnmt} = B(y_u^{Tnmt}) - C_1(y_u^{Tnmt}) - (t + \gamma) * r_u^{Tnmt} - P^u\left(r_f^{Tnmt}, y_f^{Tnmt}, r_u^{Tnmt}, y_u^{Tnmt}\right)F(y_u^{Tnmt} - r_u^{Tnmt}) \quad (10$$

The FOC for the favorite, with respect to output and misreporting decisions:

$$\frac{\partial z_f^{Tnmt}\left(y_f^{Tnmt}, r_f^{Tnmt}, y_u^{Tnmt}, r_u^{Rdm}\right)}{\partial y_f^{Tnmt}} = B_1'\left(y_f^{Tnmt}\right) - C_1'\left(y_f^{Tnmt}\right) - \frac{\partial P^f\left(r_f^{Tnmt}, y_f^{Tnmt}, r_u^{Tnmt}, y_u^{Tnmt}\right)}{\partial y_f^{Tnmt}} F\left(y_f^{Tnmt} - r_f^{Tnmt}\right) -$$

$$P^f\left(r_f^{Tnmt}, y_f^{Tnmt}, r_u^{Tnmt}, y_u^{Tnmt}\right)F'\left(y_f^{Tnmt} - r_f^{Tnmt}\right) = 0 \quad (11)$$

$$\frac{\partial z_f^{Tnmt}\left(y_f^{Tnmt}, r_f^{Tnmt}, y_u^{Tnmt}, r_u^{Rdm}\right)}{\partial r_f^{Tnmt}} = -(\mathrm{t} - \gamma) - \frac{\partial P^f\left(r_f^{Tnmt}, y_f^{Tnmt}, r_u^{Tnmt}, y_u^{Tnmt}\right)}{\partial r_f^{Tnmt}} F\left(y_f^{Tnmt} - r_f^{Tnmt}\right) +$$

$$P^f\left(r_f^{Tnmt}, y_f^{Tnmt}, r_u^{Tnmt}, y_u^{Tnmt}\right)F'\left(y_f^{Tnmt} - r_f^{Tnmt}\right) = 0 \quad (12)$$

The FOC for the underdog, with respect to output and misreporting decisions:

$$\frac{\partial z_u^{Tnmt}\left(y_f^{Tnmt},r_f^{Tnmt},y_u^{Tnmt},r_u^{Rdm}\right)}{\partial y_u^{Tnmt}} = B_1'(y_u^{Tnmt}) - C_1'(y_u^{Tnmt}) - \frac{\partial P^u\left(r_f^{Tnmt},y_f^{Tnmt},r_u^{Tnmt},y_u^{Tnmt}\right)}{\partial y_u^{Tnmt}} F(y_u^{Tnmt} - r_u^{Tnmt}) - P^u\left(r_f^{Tnmt},y_f^{Tnmt},r_u^{Tnmt},y_u^{Tnmt}\right)F'(y_u^{Tnmt} - r_u^{Tnmt}) = 0 \quad (13)$$

$$\frac{\partial z_u^{Tnmt}\left(y_f^{Tnmt},r_f^{Tnmt},y_u^{Tnmt},r_u^{Rdm}\right)}{\partial r_u^{Tnmt}} = -(\mathrm{t}+\gamma) - \frac{\partial P^u\left(r_f^{Tnmt},y_f^{Tnmt},r_u^{Tnmt},y_u^{Tnmt}\right)}{\partial r_u^{Tnmt}} F(y_u^{Tnmt} - r_u^{Tnmt}) + P^u\left(r_f^{Tnmt},y_f^{Tnmt},r_u^{Tnmt},y_u^{Tnmt}\right)F'(y_u^{Tnmt} - r_u^{Tnmt}) = 0 \quad (14)$$

As revealed by the partial derivative of audit probability with respect to reporting, each agent treats the reporting of other agents as fixed when she chooses her own reporting. This Nash conjecture leads to a Nash equilibrium. The first order conditions are not independent since agents act strategically under a tournament audit rule. In fact, the first order conditions characterize the best response functions that are solved as a system of simultaneous equations.

Since $\frac{\partial P^f\left(r_f^{Tnmt},y_f^{Tnmt},r_u^{Tnmt},y_u^{Tnmt}\right)}{\partial y_f^{Tnmt}} = -\frac{\partial P^f\left(r_f^{Tnmt},y_f^{Tnmt},r_u^{Tnmt},y_u^{Tnmt}\right)}{\partial r_f^{Tnmt}}$, the solution satisfies

$$B_1'\left(y_f^{Tnmt}\right) - C_1'\left(y_f^{Tnmt}\right) = t - \gamma \quad (15)$$

$$B_1'\left(y_u^{Tnmt}\right) - C_1'\left(y_u^{Tnmt}\right) = t + \gamma \quad (16)$$

For both the random and tournament audits, the first order conditions imply that the optimal choice of the output is independent of the reporting and of the audit scheme. Both agents will equate the net marginal gain from producing an additional unit to the per-unit cost of reporting.

The main outcomes of interest are equilibrium outputs $y_\cdot^{Rdm}$ and $y_\cdot^{Tmnt}$, and reporting decisions $r_\cdot^{Rdm}$ and $r_\cdot^{Tnmt}$. The difference between the output and reporting is misreporting, which I will compare under the random audit and tournament audits, i.e., $(y_\cdot^{Rdm} - r_\cdot^{Rdm}) - (y_\cdot^{Tnmt} - r_\cdot^{Tnmt})$. I am also interested in the how misreporting varies with the degree of heterogeneity, gamma. The specification of heterogeneity is key in this model, and I now turn to my choice of specification.

As mentioned earlier, previous studies have used both mean-preserving and non-mean-preserving specifications. Heterogeneity has been captured by differences in cost of effort (which represents ability), benefit or valuation of the prize, or contest success function. For example, Tsoulouhas & Marinakis (2007) consider heterogeneity in the ability parameter drawn from mean zero normal distribution with different variances, which results in a mean preserving heterogeneity. Ryvkin (2009) considers variation in the marginal cost of effort. Schotter & Wright (1992) introduce heterogeneity in the marginal cost of effort, as well as different wining probability

functions. In their setting heterogeneity is not mean preserving. O' Keeffe et al (1984) consider heterogeneity in marginal cost of effort, as well as different wining probability function; another type of heterogeneity that is not mean preserving. Baik (1994) also uses different valuations of the prize as well as different abilities characterized by the different cost of effort functions.

I follow Ryvkin (2013) and consider heterogeneity in the cost of reporting. The heterogeneity parameter is introduced in in such a way that it does not affect the group's average of cost of reporting. This is an important attribute of the model because it allows me to avoid a mechanical effect of heterogeneity on misreporting. In turn, this allows me to more easily isolate the behavioral channel underlying the effect of heterogeneity on reporting.

4 Results

I start by characterizing the output and reporting decisions under a random audit. Under a random audit, first order conditions have a closed form analytical solution. With the random audit, I solved the reporting and output decisions analytically. I start by characterizing the output decision across audit schemes and degrees of heterogeneity.

Proposition 1: *The optimal output level is equivalent in both the random audit and the tournament audit, regardless of agent heterogeneity.*

The intuitive explanation of this proposition is as follows. At an interior solution, agent will choose a level of output such that the marginal benefit is equal to $t + \gamma$ for the underdog and $t - \gamma$ for the favorite, under any audit framework. Given the nature of the benefit function, there is only one level of output that satisfies that condition. Therefore, the level of output is independent of reporting. Also, the average output remains constant as γ varies. Also, since marginal benefit is decreasing in output, the fact that the underdog and the favorite display marginal costs of $t + \gamma$ and $t - \gamma$ implies the following:

Proposition 2*: With heterogeneity, the favorite misreports less in equilibrium than the underdog.*

It should be noted that the difference in output chosen by the favorite and the underdog raises with the degree of heterogeneity, which results in a constant average output. Therefore, while average output is independent of the degree of heterogeneity, the dispersion of output chosen by different types of agents raises with heterogeneity.

With the tournament audit, when agents are homogenous, the optimal choices are symmetric among agents, meaning that $y_f^{Tnmt*} = y_u^{Tnmt*}$ and that $r_f^{Tnmt*} = r_u^{Tnmt*}$. As shown in proposition 2 from Cason et al (2016), the misreporting is lower with the tournament audit than with the random audit when agents are homogeneous. However, when agents are heterogeneous, due to the high nonlinearity in the probability of being audited and the asymmetric nature of output equilibrium solutions, the first-order conditions do not have a closed-form solution. To characterize the solution in this case, I resort to parametric assumptions for key benefit and cost functions, and numerically solve the first order conditions. To derive numerical solutions, I use the following parameterization and specifications shown in Table 1-1 below. The specifications are largely based on previous studies.

Table 1-1 Parameters used in numerical solution

Notation	**Definition**	**Parameters**
N	Number of players in each group	4
Number of favorites	Number of low reporting cost type	2
Number of underdogs	Number of high reporting cost type	*2*
Audit	Number of audits	2
$B(y) = \beta\, y$	Benefit of production	$\beta = 4$
$C_1(Y) = \frac{y^2}{b}$	Cost for production	$B = 59$
$C_2(r) = (t \pm \gamma) * r$	Cost for reporting	$t = 1.2;$ $\gamma = 0.2$
$F(y-r) = \frac{(y-r)^2}{f} + (t+\gamma)(y-r)$	Penalty for detected misreporting	$f = 94;$ $t = 1.2;$ $\gamma = 0.2$
$\varepsilon_i \sim U(-q, q)$	Tournament audit estimation noise	$q = 90$

The numerical results are visualized in the figures below. Figure 1-2 presents the equilibrium output and reporting decisions. First, it is straightforward to see that, as stated by Proposition 1, for a given player, the output decisions are the same regardless of the audit scheme. This is because the output choice is independent of the reporting choice. Second, the reporting,

which is the truthful revelation of output, is always higher under the tournament audit. In terms of the impact of heterogeneity, note that when agents become heterogeneous, the favorite always produces and reports more than the underdog. The dispersion between agents raises with heterogeneity.

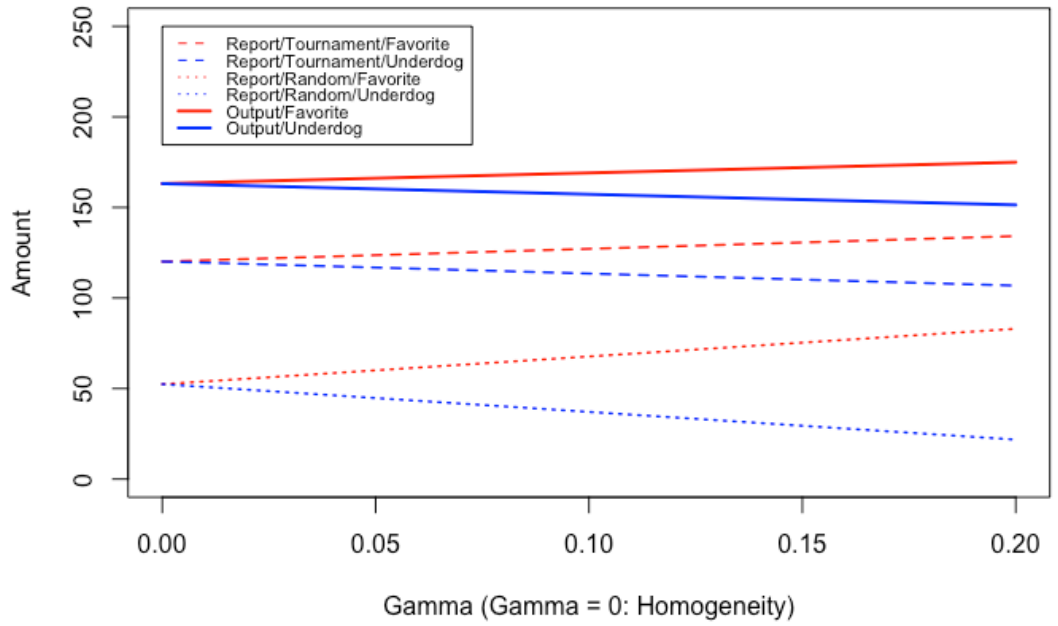


Figure 1-2 Equilibrium output and reporting

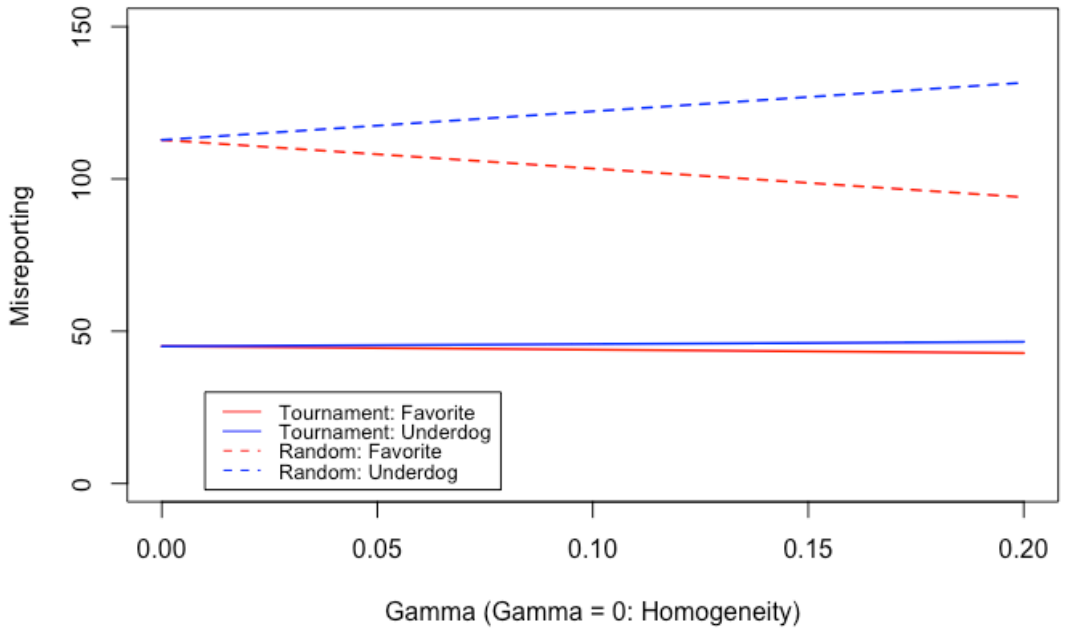


Figure 1-3 Individual misreporting by the magnitude of heterogeneity

Figure 1-3 shows individual misreporting in equilibrium. A key question we raised earlier was: what is the effect of heterogeneity on misreporting, conditional on the auditing scheme? In line with the random audit scheme, results show that, as agents become more heterogeneous, the dispersion in misreporting also grows, i.e., the underdog misreports more and the favorite misreports less. Interestingly, the increase in the underdog misreporting is proportional to the reduction in the favorite misreporting so that average misreporting remains constant.

The second question we raised was: what is the effect of the audit scheme on misreporting, conditional on heterogeneity? Our results show that a tournament auditing encourages lower misreporting on average by reducing misreporting from both players. A tournament audit also reduces the dispersion in misreporting across agents relative to the random audit, at all levels of heterogeneity. This is because the favorite's (underdog's) misreporting is incentivized (disincentivized) by the low (high) cost of reporting, but disincentivized (incentivized) by the auditing scheme.

Interestingly, the dispersion in misreporting is also much smaller in the tournament audit than in the random audit. This indicates that when agents are heterogeneous, adopting the tournament audit can also prevent extreme misreporting. Specifically, the tournament audit reduces dispersion relative to a random scheme by reducing misreporting of the underdog more than that of the favorite (in absolute and relative terms). This result is quite relevant from a policy standpoint because the underdog also tends to be the agent whose output causes more damage. Take, for instance, the case of environmental regulation. If output from an agent is more damaging, then a tax per unit of reported output will be higher. This increases the cost of truthful reporting. Therefore, the agent that causes the most damage is also the one with the highest reporting cost. In other words, underdogs are of particular concern for principals in the context of environmental regulation. With this consideration in mind, we find that a tournament audit is particularly effective in reducing misreporting by underdogs–the agents of primary concern for these regulators.

The third question we raised was: how do heterogeneity and tournament audit interact to shape misreporting? We can see that the magnitude of the reduction in average misreporting is independent of the degree of heterogeneity. This means that the effectiveness of a tournament audit to reduce misreporting found in previous studies with homogeneous agents, extend to a setting with heterogeneous agents.

To sum up the effect of the tournament audit, the results show that the tournament audit incentivizes higher truthful disclosure of the production, both in terms of the average and the individual agents especially the underdog. Importantly, the dispersion of misreporting is also substantially smaller with the tournament audit.

In terms of agents' profit, the bottom right graph in Figure 1-4 below shows the net payoff of each audit treatment in equilibrium. This shows that a tournament audit not only reduces misreporting but also raises net payoff: on average, agents' net payoff is 7% higher in the tournament audit than in the random audit scheme, both for the homogeneity and heterogeneity cases. I further decompose the total net payoff (bottom right) into net gain from production (top left), reporting cost (top right), and expected penalty (bottom left, which includes the unpaid tax and the fine). Relative to a random audit, players under a tournament audit pay a higher reporting cost, but also lower penalties. Because of the non-linearity specification of the fine function, the latter effect dominates the former, which results in agents obtaining higher net payoffs under a tournament scheme.

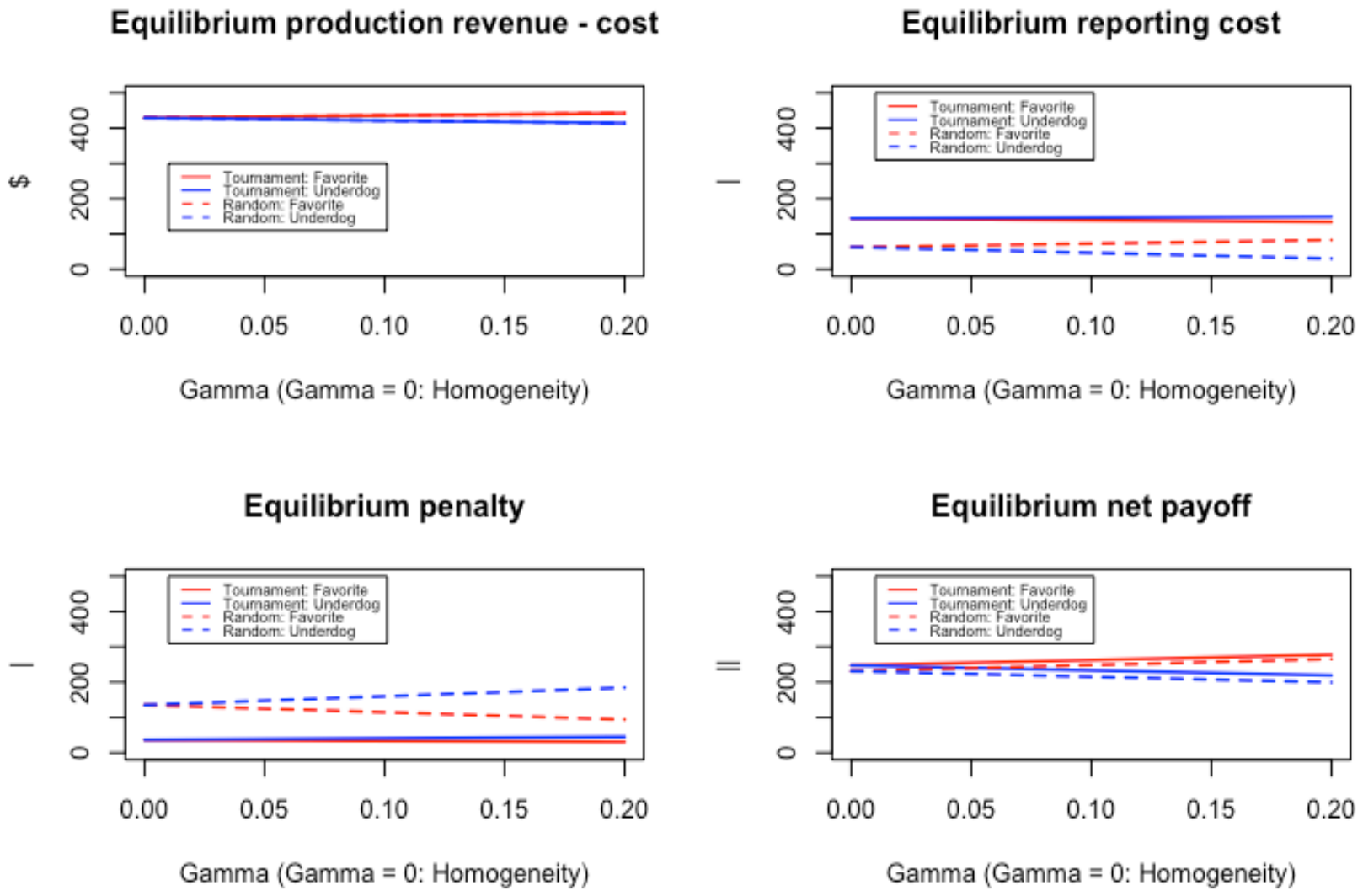


Figure 1-4 Equilibrium net payoff

We summarize our discussion in the following results:

Result 1*: With homogeneous agents, the equilibrium misreporting is lower under the tournament audit than under the random audit treatment.*

This result is a replication of Cason et al (2016), which demonstrates analytically that the tournament audit reduces misreporting with homogeneous players.

Result 2*: With heterogeneous agents, the equilibrium misreporting is lower under the tournament audit than under the random audit treatment.*

Extending from Cason et al (2016), our model shows that the advantage of tournament audit in reducing misreporting still holds even with heterogenous agents.

Result 3*: The magnitude of the reduction in misreporting attained through a tournament audit relative to a random audit is independent of the degree of heterogeneity among agents.*

This observation results from the mean-preserving spread specification of heterogeneity. As noted earlier, I model heterogeneity in a symmetric way so that average reporting cost remains the same. As a result, the change in the agent's equilibrium misreporting has the same magnitude. This intuition is visualized in the best response functions shown in Figure 1-*5*. When agents are homogeneous, the best response functions and the equilibrium are depicted by the two solid lines. When moving from tournament homogeneous to tournament heterogeneous, the favorite reduces misreporting and the underdog increases misreporting proportionally (the line across equilibrium points has a slope of -1) and the size of the change is the same, resulting in the new equilibrium on the top left of the old equilibrium.

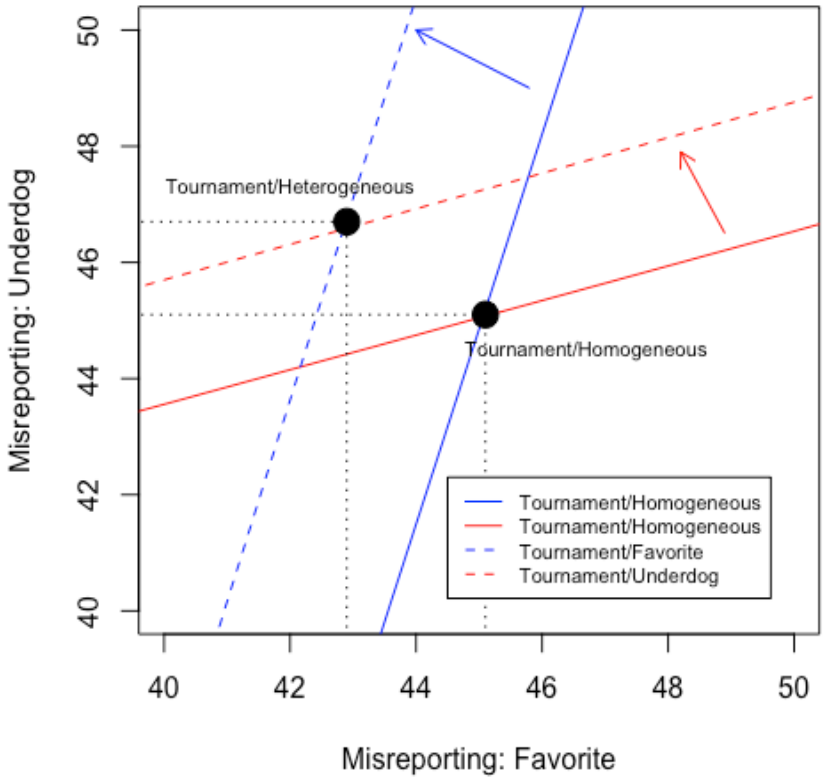


Figure 1-5 Best Response function

Result 4: *The equilibrium net payoff is higher in the tournament audit than in the random audit.*

This observation is based on the predicted net payoff shown in Figure 1-4.

Overall, the tournament audit is able to achieve much higher truthful disclosure than the random audit. When agents are homogenous, the truthful disclosure in the random audit is only 43% of that of the tournament audit. In the heterogeneity case, the truthful revelation for the favorite in the random audit is 61% of that in the tournament audit, and 20% for the underdog.

One way to think about this issue is to compute the increase in auditing frequency required under a random audit to reduce misreporting as much as a tournament audit scheme would. In fact, the theoretical model shows that to incentivize the same level of misreporting, regardless of the degree of heterogeneity, the principal would have to randomly audit 3 agents. In other words, the audit rate in a random audit has to increase from 50% (the current audit that selects two out of the four agents) to 73% (or approximately three out of four agents), to reduce misreporting to a level comparable to a tournament audit. This would, of course, increase pressure on regulatory budgets considerably. The next section discusses the robustness of these findings with respect to audit noise, the agent's composition, audit capacity, and risk aversion.

5 Robustness of main findings

5.1 Robustness to noise

The intuition behind the better performance of the tournament audit in reducing misreporting is the use of the noisy estimation of the output. This differs from why the conventional tournament can improve efficiency relative to absolute performance evaluation, such as the piece rate. First, in the conventional tournament, each agent produces an output, and the output is affected by two kinds of shocks, an idiosyncratic shock affecting only each individual, as well as a correlated common shock facing all individuals. The more important the common shock is relative to the idiosyncratic shock, the more useful such relative performance regime will be. If the output only depends on idiosyncratic shocks, then there is no information to be gained from observing the output of others. In the tournament audit, however, there is no correlated common shock. Moreover, the noisy estimation, which is similar to an idiosyncratic shock, is not utilized by the random audit. As a result, in the tournament audit, the supervisor has additional information on the true output of the agent, which gives an unbiased estimation once multiple agents are estimated.

To see how the audit noise affects the relative performance of the tournament audit as the size of heterogeneity changes, Figure 1-4 below shows the proportional difference in total misreporting at different levels of heterogeneity. The color shows the percentage difference in misreporting between random and tournament audits. For example, the red color at the bottom of the plot implies that the tournament audit can achieve a 90% reduction in total misreporting compared to the random audit. The y axis in the plot shows the audit noise. The larger the noise, the less precise the supervisor's estimation is. It can be seen that even with considerably large noise as high as 200, which is 30% more than the equilibrium output level, the tournament is still able to reduce misreporting to approximately 45% less than the random audit. In fact, as long as the audit noise is not positively infinite, the tournament audit can encourage more truthful disclosure than the random audit, though this benefit diminishes as the audit noise increases.

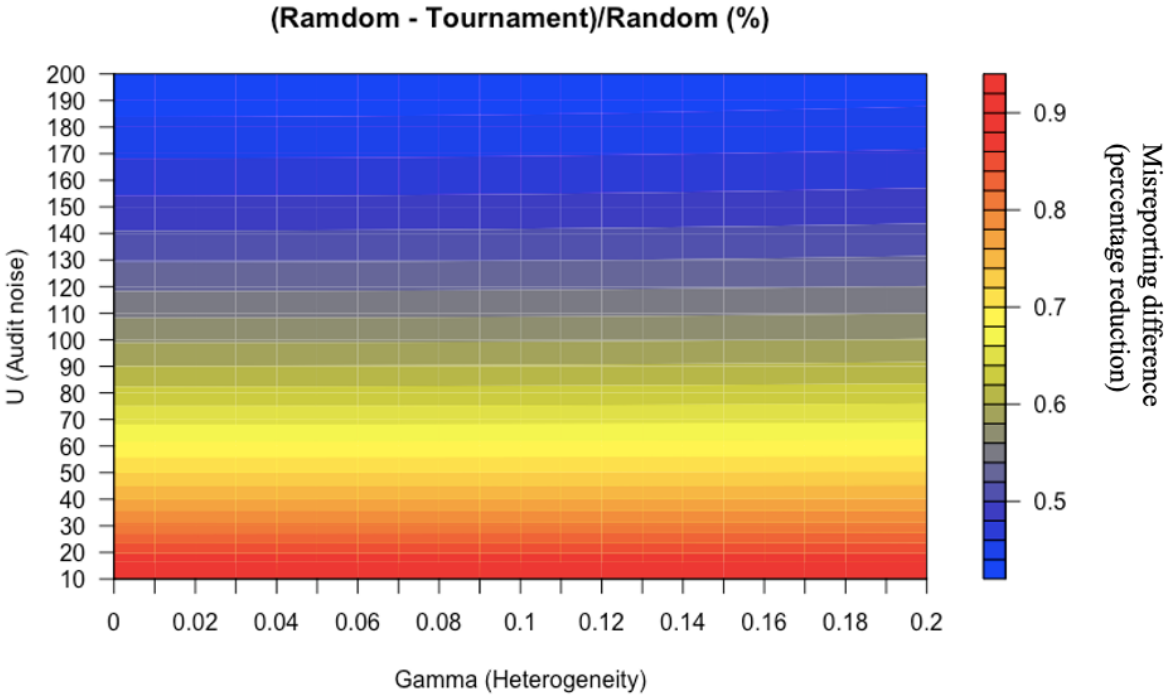


Figure 1-6 Misreporting by audit noise and heterogeneity magnitude

6.2 Robustness to composition of population

We also investigate the impact of different population compositions (that is, different compositions of favorites and underdogs in the players' pool) and audit budgets (that is, how many inspections the supervisor can conduct). In a four-player game, following Chen et al (2007), we denote different games in the following manner, shown in Table 1-2.

Table 1-2 Notations for difference games

	Number of audits		
Agents' composition	*1 Audit*	*2 Audits*	*3 Audits*
1 Favorite 3 Underdogs	1F1A	1F2A	1F3A
2 Favorites 2 Underdogs	2F1A	2F2A	2F3A
3 Favorites 1 Underdog	3F1A	3F2A	3F3A

Figure 1-7 plots the misreporting by number of audits and player composition. The results show that the composition of the population only affects the marginal probability of audit for

different players, but not the aggregate misreporting in the equilibrium. This teases out the possibility that the effect of the tournament audit, relative to the random audit we observe, is due to the even distribution of each type of player. The audit budget, on the other hand, has an impact on overall misreporting because more inspections increase the likelihood of agents being audited, lowering misreporting for both random and tournament audits.

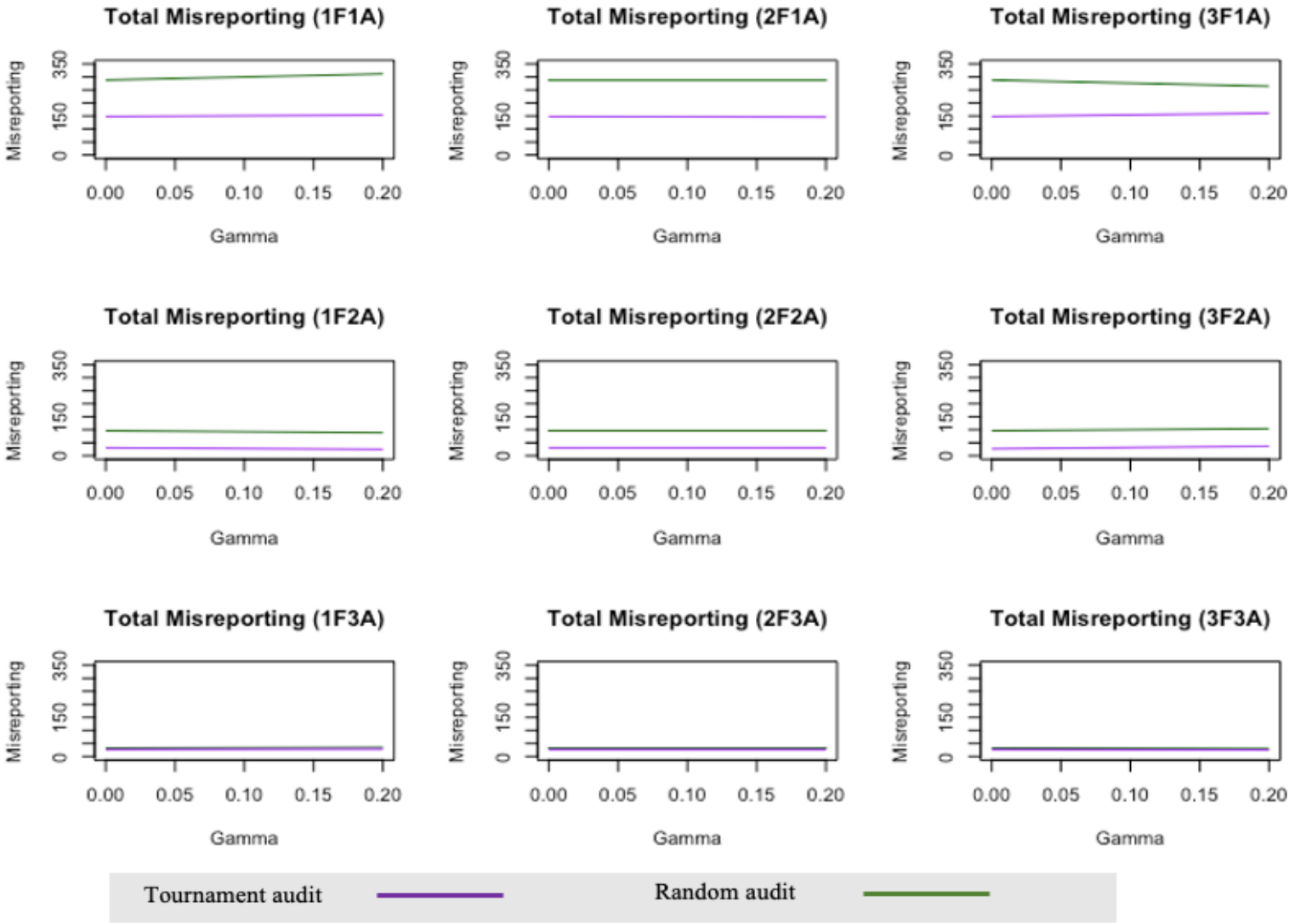


Figure 1-7 Misreporting by audit number and player composition

5.3 Robustness to risk aversion

The numerical solutions presented above are based on the assumption of risk neutrality. In this section, I discuss equilibrium and comparative statics with risk aversion. I use a constant relative risk aversion (CRRA) utility function with the coefficient of risk aversion equal to 0.5. This means the agent is rather risk averse. The numerical results are shown in Figure 1-8 below. The plot reveals that the efficiency improvement of the tournament audit is smaller when agents are risk averse. Misreporting is reduced for both random and tournament audits when agents are risk averse as opposed to risk neutral. Misreporting in a random audit, for example, decreases by around 51% when agents have a risk aversion. However, in the tournament audit, equilibrium misreporting declines much more slowly with risk aversion, with misreporting on average only falling by approximately 29%. More importantly, the tournament continues to reduce misreporting when compared to the random audit, as previously discussed.

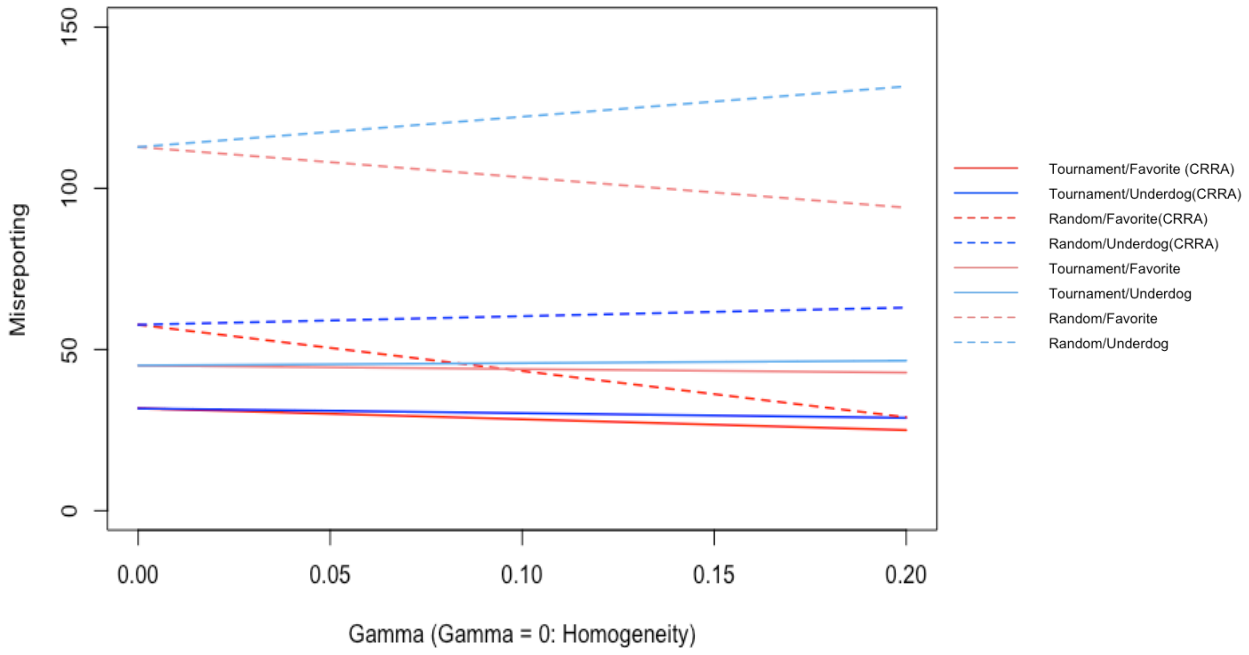


Figure 1-8 Misreporting with risk averse agents

6 Conclusion

This paper discusses the economics of information reporting under random and relative inspection standards (tournament audit). In a random audit, agents face a fixed and exogenous probability of being inspected. In a tournament audit, the relative suspicion of agents determines the chance of audit, and the agents compete with one another by reporting strategically to avoid

being audited. In this context, I am especially interested in the impact of agent heterogeneity on incentivized compliance under different audit mechanisms. I consider agents vested with different costs of reporting. Such asymmetries in reporting costs are commonly observed in practices, for example for taxpayers, polluting firms, and government agencies, but have not yet been investigated formally in the context of regulatory compliance with different audit mechanisms.

I model heterogeneity in a symmetric matter so that the average reporting cost parameter remains the same. This specification turns out to be more appropriate and realistic than the non-mean preserving specification of heterogeneity used in the contest literature. Intuitively, other things being equal, agents with lower reporting costs are expected to misreport less, so policies directed at reducing average reporting costs would be beneficial, but such policies may not always be available or can be too costly. It may, however, be feasible to make mean-preserving changes in reporting costs, for example, by transferring resources between players.

The numerical solution shows that the tournament incentives lower aggregated and individual equilibrium misreporting. This is true for different levels of heterogeneity. In the homogenous case, the level of truthful disclosure in the random audit is only 44% of that in the tournament audit. In the heterogeneity case, the truthful revelation for the favorite in the random audit is 62% of that in the tournament audit, and 20% for the underdog. The model also shows that random audit rates must be much higher to generate compliance that is comparable to what can be achieved in the tournament audit, or to significantly increase the severity of punishment. To achieve the same level of misreporting as in the tournament audit, the audit rate in the random audit would have to increase from 50% (two out of four agents audited) to 73% (three out of four agents audited), or to double the penalty parameter. The former would appear to increase the audit's cost, while the latter is not always feasible in practice. We also find that even with the same average level of misreporting, the dispersion of misreporting is substantially higher with the tournament audit. This indicates that when agents are heterogeneous, adopting the tournament audit can also prevent extreme misreporting, especially if the misreporting of the underdog is the main concern of the regulator. In addition, tournament audit also slightly improves net payoff for both all the agents regardless of the heterogeneity size by about 7%. The mechanism behind the tournament audit’s efficiency improvement is intuitive: with the random audit, the supervisor does not use the estimation of the true output, while the tournament audit relies on this estimation, which is assumed to be unbiased on average, to select which agent to audit.

The theoretical framework extends previous studies in several ways. First, I examine explicitly the relative performance of the tournament audit compared to that of the random audit as heterogeneity size changes, while the existing studies only investigate the absolute difference between tournament and random. I control the heterogeneity being incorporated in particular by using a mean-preserving spread, whereas the majority of existing studies impose heterogeneity in a non-mean-preserving manner. The theoretical results provide a framework with which researchers may reexamine the implications of heterogeneity in auditing.

The theoretical results here provide support for the strategic selection of agents (i.e., the tournament audit) by the regulatory agency or supervising department. Although the analysis is conducted for the government audit field, the model employs a fairly general formulation of the problem and could be easily adjusted for broader supervisory circumstances, such as the enforcement of environmental regulations and disclosure of firm emissions, banking regulations, individual tax compliance, and so on, where heterogeneity is prevalent and may play an important role. And the theoretical results in this paper provide a framework with which researchers may reexamine the implications of heterogeneity in auditing.

One natural question is: since the supervisor can observe the agent's type, can one create sub-audits and audits within the group of a given type of agent? It is true that, in our case, the heterogeneity is ex-ante observable, and a sorted audit is theoretically feasible. Previous studies have shown that with heterogeneous players, the favorite would prefer a pooled tournament, the underdog is indifferent, and the principal would prefer a sorted tournament (Wei, Bary, & Qin, 2019). There are two reasons I do not consider a sorted audit: first, once sorted, there is no more heterogeneity within the group and players choose the same action in the equilibrium, and even though a tournament audit still affects the marginal probability of being audit, the equilibrium audit probability converges to 50% which is exactly the same as the random audit (Ryvkin, 2011). Second, a sorted audit might not always be feasible from a regulatory perspective. Implementation and policy constraints may prohibit the sorted audit, and even with the standard tournaments, its application in the real world is usually a mix of pooled and sorted tournament so it is still relevant to study the pooled tournament audit and its effect.

To sum up, the model in this paper generates clean comparative statics depicting the effect of heterogeneity on the relative performance of audit schemes along the efficiency and surplus spectra. It is unknown to what extent these insights apply in an empirical setting. Therefore, in the

second essay, I test these predictions in a controlled environment and examine the performance of different audit schemes and the impact of players' heterogeneity using lab experiments.

7 References

Alm, J., & McKee, M. (2004). Tax compliance as a coordination game. *Journal of Economic Behavior & Organization, 54*(3), 297-312.

Alm, J., Jackson, B. R., & McKee, M. (1992). Estimating the determinants of taxpayer compliance with experimental data. *National Tax Journal*, 107-114.

Anderson, L. R., & Stafford, S. L. (2003). An experimental analysis of rent seeking under varying competitive conditions. *Public Choice, 115*(1), 199-216.

Baik, K. H. (1994). Effort levels in contests with two asymmetric players. *Southern Economic Journal*, 367-378.

Baye, M. R., Kovenock, D., & De Vries, C. G. (1993). Rigging the lobbying process: an application of the all-pay auction. *The American Economic Review, 83*(1), 289-294.

Cason, T. N., & Gangadharan, L. (2006). An experimental study of compliance and leverage in auditing and regulatory enforcement. *Economic Inquiry, 44*(2), 352-366.

Cason, T. N., Friesen, L., & Gangadharan, L. (2016). Regulatory performance of audit tournaments and compliance observability. *European Economic Review, 85*, 288-306.

Chen, K. P. (2003). Sabotage in promotion tournaments. *Journal of Law, Economics, and Organization, 19*(1), 119-140.

Clark, J., Friesen, L., & Muller, A. (2004). The good, the bad, and the regulator: An experimental test of two conditional audit schemes. *Economic Inquiry, 42*(1), 69-87.

Davis, D. D., & Reilly, R. J. (1998). Do too many cooks always spoil the stew? An experimental analysis of rent-seeking and the role of a strategic buyer. *Public Choice, 95*(1), 89-115.

Erard, B., & Feinstein, J. S. (1994). Honesty and evasion in the tax compliance game. *The RAND Journal of Economics, 1*, 1-19.

Evans, M. F., Gilpatric, S. M., & Liu, L. (2009). Regulation with direct benefits of information disclosure and imperfect monitoring. *Journal of Environmental Economics and Management, 57*(3), 284-292.

Friesen, L. (2003). Targeting enforcement to improve compliance with environmental regulations. *Journal of Environmental Economics and Management, 46*(1), 72-85.

GAO. (2019). *2019 High Risk List*. Retrieved June 19, 2019, from U.S Government Accountability Office: https://www.gao.gov/highrisk/overview

Gilpatric, S., Vossler, C. A, & McKee, M. (2011). Regulatory enforcement with competitive endogenous audit mechanisms. *The RAND Journal of Economics, 42*(2), 292-312.

Gilpatric, S., Vossler, C., & Liu, L. (2015). Using competition to stimulate regulatory compliance: a tournament-based dynamic targeting mechanism. *Journal of Economic Behavior & Organization, 119*, 182-196.

Harbring, C., Irlenbusch, B., Kräkel, M., & Selten, R. (2007). Sabotage in Asymmetric Contests– An Experimental Analysis. *International Journal of the Economics and Business, 14*, 201-223.

Harrington, W. (1988). Enforcement leverage when penalties are restricted. *Journal of Public Economics, 37*(1), 29-53.

Kräkel, M., & Schöttner, A. (2010). Technology choice, relative performance pay, and worker heterogeneity. *Journal of Economic Behavior & Organization, 76*(3), 748-758.

Lazear, P., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy, 89*(5), 841-864.

Macho-Stadler, I., & Perez-Castrillo, D. (2006). Optimal enforcement policy and firms' emissions and compliance with environmental taxes. *Journal of Environmental Economics and Management, 51*(1), 110-131.

Moldovanu, B., & Sela, A. (2006). Contest architecture. *Journal of Economic Theory, 126*(1), 70-96.

Mookherjee, D., & Png, I. (1989). Optimal auditing, insurance, and redistribution. *The Quarterly Journal of Economics, 104*(2), 399-415.

Oestreich, A. M. (2015). Firms' emissions and self-reporting under competitive audit mechanisms. *Environmental and Resource Economics, 62*(4), 949-978.

O'Keeffe, M., Viscusi, K. W., & Zeckhauser, R. (1984). Economic contests: Comparative reward schemes. *Journal of Labor Economics, 2*(1), 27-56.

Ryvkin, D. (2009). Tournaments of weakly heterogeneous players. *Journal of Public Economic Theory, 11*(5), 819-855.

Ryvkin, D. (2011). The optimal sorting of players in contests between groups. *Games and Economic Behavior, 73*(2), 564-572.

Ryvkin, D. (2013). Heterogeneity of players and aggregate effort in contests. *Journal of Economics & Management Strategy, 22*(4), 728-743.

Schotter, A., & Weigelt, K. (1992). Asymmetric tournaments, equal opportunity laws, and affirmative action: some experimental results. *The Quarterly Journal of Economics, 107*(2), 511-539.

Stowe, C. J., & Gilpatric, S. M. (2010). Cheating and enforcement in asymmetric rank-order tournaments. *Southern Economic Journal, 77*(1), 1-14.

Tsoulouhas, T., & Marinakis, K. (2007). Tournaments with ex post heterogeneous agents. Available at SSRN 1026073. *Available at SSRN 1026073*.

Tullock, G. (1980). Efficient rent seeking. In *In Efficient Rent-Seeking* (pp. 3-16). Boston, MA: Springer.

United Nations. (2014). Fundamental principles of official statistics. United Nations General Assembly.

Wei, G., Bary, B., & Qin, Y. (2019). Sorted or pooled? Optimal tournament design for heterogeneous contestants. *Cluster Computing, 22*(5), 2641-12648.

8 Appendix

Appendix A. Derivation of the tournament audit probability

This part of the appendix gives the detailed derivation of audit probability in the tournament audit. For simplicity of notation, I drop the superscript *Tnmt* because all the decision variables here are for the case of tournament audit. Without loss of generality, assuming the misreporting is greatest for the agent 1 followed by agent 2, 3, and 4 such that $\Delta_1 \geq \Delta_2 \geq \Delta_3 \geq \Delta_4$, where $\Delta_i = \mathrm{y}_i - \mathrm{r}_i$. Given the audit rule, agent i will be ranked above agent j, or $\Delta_i - \varepsilon_i \geq \Delta_j - \varepsilon_j$ if $\varepsilon_j \geq \Delta_j - \Delta_i + \varepsilon_i$, where ε is the audit noise.

Under tournament audit, the probability *i* ranked above *j* conditional on the realization of ε_i is:

$$Prob(i\ ranked\ above\ j|\varepsilon_i) = \int_{\Delta_j - \Delta_i + \varepsilon_i}^{q} \frac{1}{2q} d\varepsilon_i = \frac{P_{ij} - \varepsilon_i}{2q}$$

Where $P_{ij} = \Delta_i - \Delta_j + q$.

In this way, I can define $P_{12}, P_{13}, P_{14}, P_{21}, P_{22}, P_{23}, P_{24}, P_{31}, P_{32}, P_{34}, P_{41}, P_{42}, P_{43}$. Following Chen et al (2017), I derive the probability for each agent being audited based on different ranges of the noise ε_i. Because it is already assumed that $\Delta_1 \geq \Delta_2 \geq \Delta_3 \geq \Delta_4$, for agent 1, he will be ranked above agent 2 for sure if $\varepsilon_1 < -\Delta_2 + \Delta_1 + q = -P_{12}$. I define the range of ε_1 as Sure Audit Interval (SAI) of agent 1 relative to agent 2 if $\Delta_1 - \varepsilon_1$ is ranked above $\Delta_2 - \varepsilon_2$ for sure. Therefore, the unconditional probability that agent 1 is audited is:

$$Prob(player\ 1\ is\ audited) = \int_{-q}^{-P_{21}} \frac{1}{2q} d\varepsilon_1 + \int_{-P_{21}}^{-P_{31}} \frac{1}{2q} \frac{P_{12} - \varepsilon_i}{2q} d\varepsilon_1 + \int_{-P_{31}}^{-P_{41}} \frac{1}{2q} \frac{P_{13} - \varepsilon_1}{2q} \frac{P_{12} - \varepsilon_1}{2q} d\varepsilon_1$$
$$+ \int_{-P_{41}}^{q} \frac{1}{2q} \frac{P_{12} - \varepsilon_1}{2q} \frac{P_{13} - \varepsilon_1}{2q} \frac{P_{14} - \varepsilon_1}{2q} d\varepsilon_1$$

The first term is the probability that agent 1 is ranked above agents 2, 3, 4, for sure and ε_i is in the SAI concerning agent 2. The second term is the probability that agent 1 is ranked above agent 3, 4, for sure but ranked above agent 2 for sure, so that $\frac{P_{12} - \varepsilon_1}{2q} < 1$. The third term is the

probability that agent 1 is ranked above agent 4 for sure but not ranked above agents 2 and 3 for sure. The last term is the probability that agent 1 is not ranked above any other agents for sure.

$$Prob(player\ 2\ is\ audited) = \int_{-q}^{-P_{23}} \frac{1}{2q}\frac{P_{21}-\varepsilon_2}{2q}d\varepsilon_2 + \int_{-P_{32}}^{-P_{42}} \frac{1}{2q}\frac{P_{21}-\varepsilon_2}{2q}\frac{P_{23}-\varepsilon_2}{2q}d\varepsilon_2$$
$$+ \int_{-P_{42}}^{P_{21}} \frac{1}{2q}\frac{P_{21}-\varepsilon_2}{2q}\frac{P_{31}-\varepsilon_2}{2q}\frac{P_{41}-\varepsilon_2}{2q}d\varepsilon_2$$

$$Prob(player\ 3\ is\ audited) = \int_{-q}^{-P_{43}} \frac{1}{2q}\frac{P_{31}-\varepsilon_3}{2q}\frac{P_{32}-\varepsilon_3}{2q}d\varepsilon_3 + \int_{-43}^{P_{31}} \frac{1}{2q}\frac{P_{31}-\varepsilon_3}{2q}\frac{P_{32}-\varepsilon_3}{2q}\frac{P_{34}-\varepsilon_3}{2q}d\varepsilon_3$$

$$Prob(player\ 4\ is\ audited) = \int_{-q}^{P_{41}} \frac{1}{2q}\frac{P_{41}-\varepsilon_4}{2q}\frac{P_{43}-\varepsilon_4}{2q}\frac{P_{42}-\varepsilon_4}{2q}d\varepsilon_4$$

Without loss of generality, I assume agents 1 and 2 are the favorites, agents 3 and 4 are the underdog. Assume symmetric equilibrium within the same type of agent, the probability of being audited for each type is:

$$Prob(Favorite\ is\ audited) = Prob(player\ 2\ is\ audited) = P^f(r_f, y_f, r_u, y_u)$$
$$= \int_{-q}^{-P_{23}} \frac{1}{2q}\frac{P_{21}-\varepsilon_2}{2q}d\varepsilon_2 + \int_{-P_{42}}^{P_{21}} \frac{1}{2q}\frac{P_{21}-\varepsilon_2}{2q}\frac{P_{31}-\varepsilon_2}{2q}\frac{P_{41}-\varepsilon_2}{2q}d\varepsilon_2$$

$$Prob(Underdog\ is\ audited) = Prob(player\ 4\ is\ audited) = P^f(r_f, y_f, r_u, y_u)$$
$$= \int_{-q}^{P_{41}} \frac{1}{2q}\frac{P_{41}-\varepsilon_4}{2q}\frac{-\varepsilon_4}{2q}\frac{P_{42}-\varepsilon_4}{2q}d\varepsilon_4$$

Plugging in the expression for P_{ij}

$$P^f(r_f, y_f, r_u, y_u)$$
$$= \frac{\Delta_u - \Delta_f + 2q}{4q^2}$$
$$+ \frac{[(\Delta_u - \Delta_f + 2q)4q^2 + 2q(\Delta_f - \Delta_u + q)(\Delta_u - \Delta_f + 2q) + (\Delta_u - \Delta_f - 2q)(\Delta_u - \Delta_f + 2q) + \frac{2q^2}{3} - (\Delta_u - \Delta_f + 2q)^3]}{16q^4}$$

$$P^u(r_f, y_f, r_u, y_u)$$
$$= \frac{\Delta_u - \Delta_f + 2q}{4q^2}$$
$$+ \frac{[(\Delta_u - \Delta_f + 2q)4q^2 + 2q(\Delta_f - \Delta_u + q)(\Delta_u - \Delta_f + 2q) + (\Delta_u - \Delta_f - 2q)(\Delta_u - \Delta_f + 2q) + \frac{2q^2}{3} - (\Delta_u - \Delta_f + 2q)^3]}{16q^4}$$

I then use these expressions to solve for the misreporting numerically. The code used is shown in the next section of the appendix.

Appendix B. Code for numerical results

```
require("nleqslv")
options(scipen = 999)
options(digits=3)
beta  <- 4
a     <- 90
t     <- 1.2
b1 <- 59
b2 <- 94
b22 <- 2*b2
g <- 0.2
Gamma <- seq(0, g, by=g)
#################### 2f2u ####################
f_tnmt_r    <-  matrix(nrow=length(Gamma), ncol=1)
u_tnmt_r <-  matrix(nrow=length(Gamma), ncol=1)
f_rdm_r     <-  matrix(nrow=length(Gamma), ncol=1)
u_rdm_r  <-  matrix(nrow=length(Gamma), ncol=1)
f_tnmt_y    <-  matrix(nrow=length(Gamma), ncol=1)
u_tnmt_y <-  matrix(nrow=length(Gamma), ncol=1)
f_rdm_y     <-  matrix(nrow=length(Gamma), ncol=1)
u_rdm_y  <-  matrix(nrow=length(Gamma), ncol=1)
for (i in 1:length(Gamma)){
gamma   <- Gamma[i]
fn_tnmt <- function(x) {
rf    <- x[1]
ru    <- x[2]
yf    <- x[3]
yu    <- x[4]
```

```
p31  <- yu-ru - (yf-rf) + a
p13  <- yf-rf - (yu-ru) + a
p32  <- p31
p23  <- p13
p24  <- p23
p42  <- p32
p14  <- p13
p41  <- p31
p12  <- a
p21  <- a
p34  <- a
p43  <- a
t1   <- (1/(2*a)) * (yf-rf - (yu-ru))
f3   <- function(epsilon) {(1/(2*a))* ((p12+epsilon)*(p13+epsilon)/((2*a)^2) +
(p12+epsilon)*(p14+epsilon)/((2*a)^2)+(p13+epsilon)*(p14+epsilon)/((2*a)^2)-
2*(p12+epsilon)*(p13+epsilon)*(p14+epsilon)/((2*a)^3))}
t3   <- integrate(f3, lower = -a, upper = p41)
f5          <- function(epsilon) {(1/(2*a))* ((p41+epsilon)*(p42 + epsilon)/((2*a)^2) +
(p41+epsilon)*(p43 + epsilon)/((2*a)^2) +        (p42+epsilon)*(p43 + epsilon)/((2*a)^2) -
2*(p41+epsilon)*(p42 + epsilon)*(p43 + epsilon)/((2*a)^3))}
t5   <- integrate(f5, lower = -p41, upper = a)
probf  <- t1 + t3$value
probu  <- t5$value
mp_fwin <- 1/(2*a) - 1/((2*a)^3)
mp_uwin <- 1/(2*a) - 1/((2*a)^3)
ef_y  <- (beta) - (yf)/b1 - probf*((yf-rf)/b2 + t-gamma) - mp_fwin*((yf-rf)^2/b22+(t-gamma)*(yf-
rf))
eu_y          <- (beta) - (yu)/b1 - probu*((yu-ru)/b2 + t+gamma) - mp_uwin*((yu-
ru)^2/b22+(t+gamma)*(yu-ru))
ef_r  <-  -t+gamma + probf*((yf-rf)/b2+t-gamma) + mp_fwin *((yf-rf)^2/b22+(t-gamma)*(yf-rf))
```

```
eu_r <- -t-gamma + probu*((yu-ru)/b2+t+gamma) + mp_uwin *((yu-ru)^2/b22+(t+gamma)*(yu-ru))
return(c(ef_y, eu_y, ef_r, eu_r))}
result_tnmt <- nleqslv(c(15,15,15,15), fn_tnmt)
x_tnmt <- result_tnmt$x
xstar_tnmt <- ifelse(x_tnmt > 0, x_tnmt, 0)
fn_rdm <- function(x) {
rf <- x[1]
ru <- x[2]
yf <- x[3]
yu <- x[4]
ef_y <- (beta) - (yf)/b1 - 0.5*(((yf-rf)/b2) + t-gamma)
eu_y <- (beta) - (yu)/b1 - 0.5*(((yu-ru)/b2) + t+gamma)
ef_r <- -t+gamma + 0.5*((yf-rf)/b2+t-gamma)
eu_r <- -t-gamma + 0.5*((yu-ru)/b2+t+gamma)
return(c(ef_y, eu_y, ef_r, eu_r))}
result_rdm <- nleqslv(c(15,15,15,15), fn_rdm)
x_rdm <- result_rdm$x
xstar_rdm <- ifelse(x_rdm > 0, x_rdm, 0)
f_tnmt_r[i,1] <- xstar_tnmt[1]
f_rdm_r[i, 1] <- xstar_rdm[1]
u_tnmt_r[i,1] <- xstar_tnmt[2]
u_rdm_r[i, 1] <- xstar_rdm[2]
f_tnmt_y[i,1] <- xstar_tnmt[3]
f_rdm_y[i, 1] <- xstar_rdm[3]
u_tnmt_y[i,1] <- xstar_tnmt[4]
u_rdm_y[i, 1] <- xstar_rdm[4]}
```

ESSAY 2. STRATEGIC MISREPORTING UNDER ALTERNATIVE AUDIT MECHANISMS WITH HETEROGENEOUS AGENTS: EXPERIMENTAL EVIDENCE

Abstract

This paper examines the effectiveness of different regulatory schemes for reducing misreporting when agents are heterogeneous. We conduct experiments where we compare two stochastic auditing schemes: random audit whereby agents are randomly chosen for inspection, and tournament audit where the probability of inspection raises with the agent's estimated underreporting. To examine the role of heterogeneity, the experiment varies the cost of reporting (or, conversely, the benefit of underreporting) across agents. Our experimental results are largely consistent with theoretical predictions from Essay 1. We find that output is independent of the auditing mechanism, and that a tournament scheme reduces average misreporting. More importantly, the reduction in average misreporting is independent of the degree of heterogeneity among agents. This means that findings from previous experiments that examined the effects of tournament audits under homogeneous agents can be generalized to a setting with heterogeneous agents. However, the average agent misreports less than predicted by theory, and the net payoff is not significantly higher in the tournament audit, unlike what is predicted. These are relevant deviations from theoretical predictions and might be partially explained by the risk aversion. The results imply that the tournament audit improves efficiency relative to the random audit, but such gains might be smaller than theoretically predicted, and that a tournament might not be as effective in improving surplus as the model predicts. This shows that the main insights obtained with homogeneous agents (Cason et al., 2016) are generalizable to a setting with heterogeneous agents, as long as heterogeneity is mean preserving, i.e., as long as the average cost of reporting remains constant at varying levels of heterogeneity.

1 Introduction

Many regulatory settings are characterized by agents that have private information regarding their performance (henceforth, output) and must report to a principal. To effectively regulate output, the principal must first design and implement a scheme to reduce misreporting by

the agent. Some prominent schemes rely on auditing with penalties for misreporting. Previous studies (e.g., Cason et al., 2016) have compared the effectiveness of these schemes, but in an environment where all agents have the same cost of reporting (or, conversely, benefit from misreporting).

In this paper, we examine the effectiveness of these auditing schemes with heterogeneous agents, i.e., agents that differ in their cost of reporting. This is important because heterogeneity among agents is not only prevalent in the context of, for example, compliance with environmental regulations, but it has also been shown to influence the effectiveness of regulatory schemes in other settings. For instance, in the conventional tournament literature, heterogeneity lessens the effectiveness of a scheme based on relative performance (tournaments) vis-à-vis a scheme based on absolute performance (piece-rate). It is unclear whether and to what extent these insights translate to auditing schemes. This is why it is crucial to better understand how heterogeneity shapes the relative effectiveness of alternative auditing schemes, as well as their efficiency and distributional implications.

If a principal can audit an agent, then they are able to observe their output. But audits are costly, and budgets are limited, so only a fraction of the agents will be audited by the principal. Nevertheless, misreporting can be discouraged by the probability of being audited in combination with a penalty if the audit does uncover misreporting. The probability that an agent is inspected is determined by the auditing scheme. In one prominent scheme, called random audit, agents are randomly chosen for auditing. In another prominent scheme, called tournament audit, the principal has a noisy estimate of output which, in combination with reported output, results in a noisy estimate of misreporting. Agents whose misreporting is relatively high, are more likely to be audited by the principal. Previous literature has found that a tournament scheme reduces misreporting when agents are homogeneous (Gilpatric et al., 2011; Gilpatric et al., 2015; Cason et al., 2016). Our objective is to examine whether this key result generalizes to settings where agents are heterogeneous. Theoretical predictions from the first essay suggest it does. In this study we examine experimental evidence to test this. We also study the effect of tournaments on misreporting along the reporting cost spectrum.

Empirically examining the effect of alternative audit schemes on misreporting using observational data is challenging because data on actual misreporting is seldom available, and quasi-experiments are fraught with selection issues, making it hard to establish causality. Lab

experiments, on the other hand, allow us to overcome both of these problems aiding causal inference. The experiment randomly assigns treatments, generating more comparable treatment and control groups. Moreover, an experimental investigation confers us with stricter control of the environment. We exploit this to implement a strategic environment that closely resembles our theoretical framework, thereby facilitating comparisons between experimental results and theoretical predictions (obtained from Essay 1). We randomly select subjects into one of four treatments: random audit with homogeneous agents, random audit with heterogeneous agents (where an agent can have a high or low cost of reporting), tournament audit with homogeneous agents, and tournament audit with heterogeneous agents. All sessions were conducted in Spring 2021 at Purdue University.

The experimental evidence shows that agents in the lab lie less than predicted by the theory. As a result, the magnitude of the effect of the tournament audit is less than predicted and the net payoffs are lower than what the agent could receive if misreported "optimally". However, the tournament audit still significantly reduces misreporting compared to the random audit, thus supporting, at least partially, the comparative statics generated by the theoretical model. In particular, we find that a tournament audit reduces misreporting under heterogeneity, which indicates that results from Cason et al (2016) are generalizable to an environment with heterogeneous agents. This also stands in contrast with the standard tournament literature, where heterogeneity reduces the effectiveness of tournaments relative to a piece-rate scheme. The learning process is significant, such that the subjects' decisions during the later rounds are much closer to the equilibrium predictions. Perhaps more importantly, under heterogeneity, the difference in misreporting between low- and high-cost agents in a random audit experiment is small. This means that a switch from a random to a tournament scheme reduces misreporting from low- and high-cost agents proportionally. Moreover, as predicted by theory, the quantitative effect of tournament audits on misreporting is independent of the degree of heterogeneity.

There are a few ways in which experimental results differ from theoretical predictions. First, subject output choices are about 30% lower than predicted, despite the fact that it is the same across audit schemes, as predicted by theory. Also, under a random audit scheme, agents misreport much less than theory predicts. Moreover, the average net payoffs between tournament and random audit treatment are not significantly different, and due to both the under-production and the under-misreporting, an average player in the experiment owns around 7 dollars from the game, while the

theoretical model predicts an equilibrium net payoff should be around 20 dollars. In sum, the experimental evidence indicates that a tournament audit is preferred to a random audit regardless of the degree of the heterogeneity but that the efficieny improvement of a tournament audit are more limited than theory would suggest. Since most of the divergence from theoretical predictions happens under the random audit treatment, it might be partially explained by the risk aversion of agents, which has a much larger impact on decision making under the random than under the tournament, as suggested in Essay 1, although the risk aversion does not rationalize the output level chosen in the experiment.

This paper is most closely related to previous experimental studies investigating misreporting under random and tournament audit schemes. These studies assume homogeneous agents and have found that a tournament audit reduces misreporting compared to a random audit (e.g., Gilpatric et al., 2011; Gilpatric et al., 2015; Cason et al., 2016). However, in most regulatory settings, agents are different in their ability to hide information, i.e., the reporting cost for different agents is inherently heterogeneous. For example, in the public sector, some departments might face more drastic sanctions for misreporting, or may find it more difficult to misreport. Similarly, among private agents, governments usually deem pollution from some agents deserves harsher penalties than the same pollution from other agents (e.g., runoff from farms located close to waterways). Our experiments, therefore, build on previous work and expands our understanding of the effect of alternative audit schemes to more realistic settings in which agents are heterogenous. Our analysis shows that, the ability of tournament audits to reduce misreporting extends to regulatory settings with heterogeneous agents, but that the efficiency gains may be more limited than theoretically predicted.

Our insights are applicable to a relatively broad set of regulatory circumstances where heterogeneity is widely present, and many play an important role in affecting the agent's decisions. These include enforcement of environmental regulations and disclosure of firm emissions, banking regulations, and individual tax compliance, among others.

2 Literature Review

The structure of the contest has been applied to studying competitive endogenous audits, or tournament audits. A tournament audit is distinguished from a regular tournament in at least two ways. First, the level of misreporting not only affects who gets audited, but the penalty is

conditional upon being audited. This differs from the standard tournament in which the effort only affects the chance of winning but not the amount of the prize. Second, the audit noise in the tournament audit is similar to the idiosyncratic shock in the standard tournament, which is not relevant under the random audit mechanism.

Several experiments have found evidence that contest-based audit schemes reach higher levels of disclosure relative to random audits. Alm et al (1992) studied the effect of different tax audit rules in the lab and found that the cut-off rule–where reported income above a certain level will be audited–is the most effective scheme. Gilpatric et al (2011) found that conditional audit schemes generate higher truthful disclosure. Stowe & Gilpatric (2010) tested cheating in a tournament under a correlated random audit. Cason et al (2016) also focused on disclosure decisions of agents under different audit mechanisms. They found that endogenous audit mechanisms incentivize better compliance relative to random audits. Their results also show that misreporting is relatively stable over time under the endogenous audit but increases considerably with experience with random auditing[4]. One of the common features in the above-mentioned studies is homogeneous agents.

While many studies set output as an exogenous signal, Oestreich (2015) endogenizes output and reporting. He models both the Tullock contest (1980) and all-pay auction and finds that both strategies increase truthful reporting than the random audit. Evans et al (2009) model enforcement as the choice of the regulator who faces a limited budget and finds the tradeoff between truthful reporting and emissions. This strand of study, though, assumes symmetric agents.

Overall, agents' heterogeneity has been studied extensively in standard contests, but not in tournament audits. On the other hand, as summarized in Essay 1, the reporting behavior of identical agents in an endogenous audit game is well understood theoretically. Direct empirical evidence of the impact of heterogeneity is scarce in the context of the tournament audit mechanism. I am aware of only two published economic experiments that consider the effects of endogenous audit rules

[4] This has also been tested in a dynamic setting where the audit probability can be conditioned on past compliance (See Cason & Gangadharan (2006) and Harrington (1988) for example). Clark et al (2004) test the compliance rate in lab under two dynamic targeting schemes: the Harrington's (1988) and Friesen's (2003) that differs in the Marcov transition rules and find that although both strategy lower the inspection rate, both reaches lower compliance rate than random audit. Gilpatric et al (2015) also study the property of tournament-based audit in a dynamic game.

with somewhat heterogeneous agents: Gilpatric et al (2011) and Gilpatric et al (2015) examine heterogeneity in firms' emissions. The emissions are exogenously given and differ across firms, reflecting the fact that even with similar production technology, firms' actual emissions might differ. Nonetheless, firms remain strategically symmetric competitors, and their setting does not reflect the fact that firms differ ex-ante.

A novel feature of this paper is the incorporation of cost heterogeneity in tournament audits. Considering the impact of heterogeneity in the context of a government audit is perhaps more relevant empirically because of the natural occurrence of bureaucratic diversity. For example, some bureaucracies have stronger administrative influence over statistical departments, thus bearing the lower cost of manipulating statistics. Relaxing the symmetry assumption, therefore, permits a somewhat richer version of the government structure and reporting behavior of different reporting agents. The following section introduces an experimental design as well as testable hypotheses.

3 Experimental designs

3.1 Decision making

The experimental design is a 2 by 2 factorial design, which is displayed in Table 2-1. I use a between-subjects design such that each subject was only exposed to one treatment. In each treatment, subjects make decisions in 24 rounds. All the parameters used in the experiment are common knowledge. Decision making and the environment, the matching protocol and role switching rule will soon be described in more detail.

Table 2-1 Treatments and cell design

Treatment	*Random audit*	*Tournament audit*
Homogeneous agents	24 subjects 24 rounds	24 subjects 24 rounds
Heterogeneous agents	24 subjects 24 rounds	24 subjects 24 rounds

In each of the 24 rounds, the subjects make two decisions. They first decide how much to produce by entering a number on their computer screen. Subjects can choose a level of output between 0 and 300. Output is associated with per unit monetary private payoff which is same for all subjects and a convex private production cost. The second decision is to choose what level of

output to report to the supervisor, the computer. For each unit of output reported, subjects pay a reporting cost. This creates an incentive for underreporting. In a setting with homogeneous agents, this reporting cost is the same for all subjects. In a setting with heterogeneous agents, the reporting cost is lower for the favorite. Although there is no incentive to overreport the output, in the experiment the subject can report any number they like, which can be greater than, less than or equal to their output choice.

Once all choices are made, the supervisor (computer) chooses two out of the four subjects to audit. Subjects were informed that, if audited, the audit process will reveal their actual output and, consequently, the magnitude of misreporting, if any. If the subject has misreported and is audited, there is a penalty (the term "additional cost" was used in the experiment) based on the magnitude of the misreporting. The penalty, as described in Essay 1, includes the unpaid reporting cost plus a convex fine. The subject does not know the output and reporting decisions on another subject's screen.

In each round, two of the four subjects in each group will be audited. There are two audit rules: random and tournament audit. Under the random audit, two out of the four subjects are randomly selected to be audited. Under the tournament audit, the subject is informed that his/her misreporting and the paired subjects' misreporting jointly determine his/her audit probability. Specifically, after all four subjects make reporting decisions, the computer will rank the four subjects based on the estimated output minus how much they report. The noise in the estimation has an equal chance of being any number between$[-90,90]$, and on average, the estimation equals the actual amount. The two subjects with the greatest difference between the estimated and the reported number will be inspected, and ties will be broken randomly by flipping a coin. The other two subjects will not be audited in that round.

Table 2-2 shows an illustrative example of how the audit is determined in a certain round. In the experiment, the subjects were told that the computer observes each subject's actual output with a random amount (noise), which gives estimated misreporting shown in the last column, and the computer audits the two players with larger estimated misreporting, in the illustrative example, subject 2 and 4 (even though they may not be the ones that actually misreported the most). Notice that since the audit noise is between -90 and 90, if a subject chooses to misreport more than 90, he or she will be audited for sure; this is the "sure audit interval" as mentioned in Essay 1.

Table 2-2 An illustrative example of how audit is decided in the tournament treatment

	Actual Output	Reporting	Output estimated by the computer	Gap estimated by the computer	Inspection
	A	**R**	**A + *random amount***	**A + random amount - R**	
Member 1	100	95	100 - *11* = 89	89 – 95 = -6	No
Member 2	200	200	200 + *63* = 263	263 – 200 = **63**	**Yes**
Member 3	152	111	152 – *39* = 113	113 – 111 = 2	No
Member 4	80	70	80 + *25* = 105	105 - 70 = **35**	**Yes**

3.2 Heterogeneity

The heterogeneity is captured via different per unit reporting costs. As specified in Essay 1, I use a monetary reporting cost function that specifies each agent's cost of reporting as how much money they spend at each reporting level.[5] During the experiment, for the homogeneity treatment, each subject was shown the zTree screenshot that shows "you pay 1 experimental dollar for each unit of output reported". In the heterogeneity treatment group, each subject was shown two zTree screens, one shows "You are a Low-cost type in this round, you pay 1 experimental dollar for each unit of output reported." And the other shows "You are a High-cost type in this round, you pay 1.4 experimental dollars for each unit of output reported."

Notice that for the heterogeneity treatment, the favorite/underdog type is not fixed and the role switching takes place in blocks. Role switching helps the subject to put him or herself in the shoes of the other agents and is especially useful for learning. In the experiment, the role is randomly assigned in blocks of 12 rounds. The subjects were specifically informed that their reporting costs could be high or low. The subjects were also informed that they would be assigned to one of two types for the first 12 rounds and the other type for the second 12 rounds, with an equal chance of success in each. For example, they can be an underdog (or, as said in the instruction, the "high-cost type") for the first 12 rounds and a favorite ("low-cost type") for the second 12 rounds. Or with an equal likelihood they can be a favorite ("low-cost type") for the first 12 rounds and an underdog ("high-cost type") for the second 12 rounds. The groups were then rematched using stranger matching after each round.

[5] The monetary cost function has been used in various contests experiments, for example Bull & Schotter (1987).

All lab sessions were conducted at the Vernon Smith Experimental Economics Lab at Purdue University using Z-tree software (Fischbacher, 2007). The participants were undergraduate and graduate students at Purdue University with diverse majors and backgrounds. Sessions were conducted during the spring semester of 2021. There were eight sessions, and there were 12 subjects participated in each session. Recruitment of the subjects was done by ORSEE (Greiner, 2004). Each subject can participate in one experimental session only. Subjects were paid based on how much they earned in the experiment, plus a show-up fee. Besides the amount they earn by making output and reporting decisions, at the end of the instructions, subjects take a quiz to examine their understanding of the game rules. Subjects were paid based on the number of correctly answered questions. If the subject fails to answer any question correctly, the correct answer shows up on his/her screen to help them understand the instruction.

After the experiment, demographic information like gender and age was collected in order to be used as additional control variables. Moreover, the subjects also participated in an investment task to elicit their risk aversion. For this task, each subject received a $5 endowment and had the option to invest as much as they wanted. The investment has an equal opportunity to return either zero or three times the amount invested (Gneezy and Potters, 1997). The final payoff for each subject was the total net earnings in 8 randomly chosen rounds plus a 7-dollar show-up fee that is the same for all subjects, the quiz earnings, and the investment task. The average total earned in the experiment was $31.4, with a range of $14 to $42.5. Sessions usually last about 60 minutes on average, including the time taken for reading instructions at the beginning and payment distribution at the end of the experiment.

To minimize potential emotional aversion to misreporting decisions, during the experiment I used neutral framing in describing the decision-making environment. I used the terminology of output, reporting, the gap between output and reporting, inspections, additional costs consistently throughout all four treatments. In particular, the term “misreporting” was not mentioned throughout the experiment. Rather, I use expressions such as: “Your decision is to choose what output number to report to the computer. For each unit of output reported that exceeds the number you see, you pay a reporting cost. This cost is deducted from your earnings. You can choose to report any amount you like, such as your actual output or an amount less or more”. All the parameters and specifications used in the experiment are summarized in Essay 1 Table 1-1.

3.3 Testable hypotheses

Based on the theoretical predictions from Essay 1, I test several sets of hypotheses. I then compare the experimental results with these theoretical predictions to identify which parts of our theory can actually be falsified, and what is the source of the deviation between theoretical outcomes and their empirical analog. There are four sets of main hypotheses. Notice that although the predictions from the theoretical model deal with one-shot instead of repeated games, the experiments were conducted in multiple, repeated rounds. This is because subjects face a complex decision-making task, the output and reporting decisions in the first few rounds might be driven by the fact that they have not yet fully understood the task. Moreover, since there are no reputational effects, the only SPNE involves the choice of Nash equilibrium misreporting levels for the one-shot game in each round. Table 2-3 presents the equilibrium predictions by audit treatment and by heterogeneity treatment, as well as by different agent types.

Table 2-3 Equilibrium predictions

Audit	Homogeneity/ Heterogeneity	Agent type	Output	Reporting	Misreporting	Misreporting as % of output	Net Payoff
Random	*Homogeneity*	/	161	51	110	67%	231
	Heterogeneity	Favorite	174	82	92	52%	265
		Underdog	148	20	128	85%	199
Tournament	*Homogeneity*	/	161	120	41	24%	249
	Heterogeneity	Favorite	174	134	40	22%	278
		Underdog	148	106	42	27%	219

The first set of hypotheses is based on Proposition 1 from Essay 1 and is as follows:

Hypothesis 1: output

Hypothesis 1a: With the homogeneity treatment, the output level chosen is the same in the random audit and tournament audit treatments.

Hypothesis 1b: With the heterogeneity treatment, the output level chosen is the same in the random audit and tournament audit treatments.

The second set of Hypothesis is based on Results 1 to 3 from Essay 1:

Hypothesis 2: misreport

Hypothesis 2a: With the homogeneity treatment, tournament audit treatment reduces misreport than that in the random audit treatment.

Hypothesis 2b: With the heterogeneity treatment, tournament audit reduces misreport than the random audit treatment.

Hypothesis 2c: the difference in misreporting between tournament and random audit remains the same

The third set of hypotheses is based on Proposition 2 from Essay 1

Hypothesis 3: favorite versus underdog

Hypothesis 3a: the favorite misreports less than the underdog regardless of the audit treatment.

Hypothesis 3b: With heterogeneity treatment, both the underdog and the favorite's misreport less in the tournament audit.

Hypothesis 3c: In the random audit treatment, the difference in misreporting between the favorite and the underdog is greater than in the tournament audit treatment.

Inspired by Results 4 from Essay 1, Hypothesis 4 is related to the net payoff in each audit treatment.

Hypothesis 4: Net payoff

The net payoff is greater in the tournament audit treatment than in the random audit treatment.

4 Results

I ran 8 experimental sessions with 96 subjects. Gender, race, major and academic performance are well balanced across the treatments. T-tests show that none of the demographic characteristics are significantly different across treatments. The average age of the subjects is 21 years old. Around 25% of the subjects are from the Management or Business background. An average subject has a GPA between 3.5 and 4, currently in the third or fourth year of college, and has participated in at least one other economic experiment (not this particular experiment) before. For the investment task, 96% of the subjects invested some money out of their 5-dollar endowment. The bimodal investment appears on 2.5 (invests half of the endowment) and 5 (invests all the endowment). For more detail about the demographics and investment decisions of the subject, see the summary statistics in the appendix.

Overall, I find that the experiment results support the theoretical predictions of the effect of the tournament audit. Although the absolute level of misreporting is below the theoretical prediction under all treatments. First, the tournament audit always leads to lower misreporting compared to the random audit, regardless of heterogeneity. Such differences are largely due to

differences in the amount of underreporting rather than differences in the proportion of subjects who underreport. The difference in misreporting between the tournament and the random audit is unaffected by heterogeneity (difference in difference =0). I also find that, consistent with the theory, the favorite always misreports less than the underdog in both audit treatments towards the later rounds of the game, as subjects learn the game. The difference in misreporting between the favorite and the underdog is smaller under the tournament audit, although such different is not significant. Finally, although the theoretical model predicts a slightly higher net payoff under the tournament audit, in the experiment, such a higher payoff is not significant.

Table 2-4 Decisions by treatment

Average of all rounds						
Audit	**Homogeneity/Heterogeneity**	**Agent type**	**Output**	**Report**	**Misreport**	**Misreport (% of output)**
Random	*Homogeneity*	/	126.2 (14.9)	81.0 (18.1)	45.2 (8.7)	36% (8%)
	Heterogeneity	Favorite	131.5 (19.2)	79.1 (22.9)	52.4 (12.3)	39% (7%)
		Underdog	116.8 (8.7)	65.4 (17.5)	51.3 (16.8)	42% (1%)
Tournament	*Homogeneity*	/	120.1 (15.5)	92.67 (18.1)	27.3 (4.1)	24% (5%)
	Heterogeneity	Favorite	125.8 (12.3)	95.7 (15.9)	30.1 (10.2)	25% (8%)
		Underdog	119.6 (16.5)	89.64 (16.9)	29.9 (5.3)	26% (4%)

Average of the last five rounds						
Audit	**Homogeneity/Heterogeneity**	**Agent type**	**Output**	**Report**	**Misreport**	**Misreport (% of output)**
Random	*Homogeneity*	/	120.0 (4.6)	64.5 (4.8)	55.4 (2.2)	47% (2%)
	Heterogeneity	Favorite	128.1 (9.3)	69.5 (9.1)	58.5 (6.9)	45% (4%)
		Underdog	123.2 (0.8)	49.4 (5.0)	73.6 (5.5)	57% (3%)
Tournament	*Homogeneity*	/	112 (1.4)	80.7 (1.8)	31.3 (1.2)	31% (1%)
	Heterogeneity	Favorite	122.9 (7.6)	81.2 (6.0)	41.8 (8.2)	34% (3%)
		Underdog	116.3 (5.8)	80.78 (3.3)	35.5 (3.1)	30% (1%)

Notes: average of all rounds (top table) vs last five rounds (bottom table). Standard deviations are in paratheses.

Table 2-4 shows the average decision by treatments. I pool the decisions of all subjects in all rounds on the top table and I pool the decisions in the last five rounds on the bottom table, separately. First, the output decisions tend to be lower in the last five rounds, while the misreporting decisions tend to be higher in the last five rounds. Both the favorites and underdogs misreport less when facing the tournament audit relative to the random audit, especially in the last five rounds. The standard deviations for both output and misreporting decisions are also smaller in the last five rounds. These results suggest learning on the part of the subjects. We now turn to this issue.

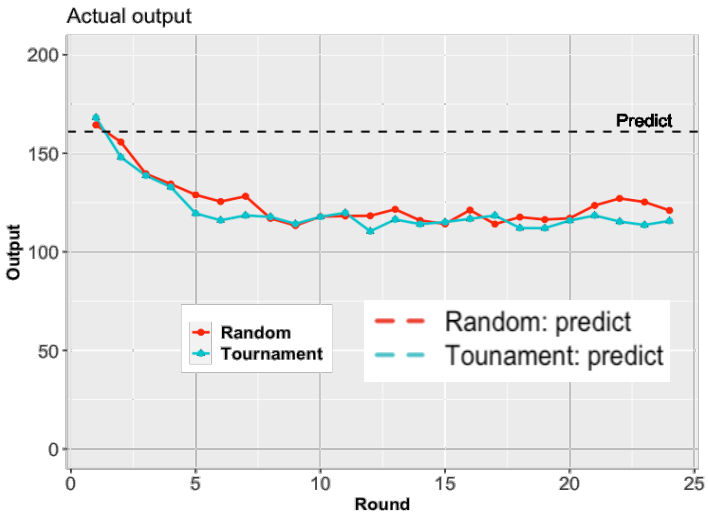
4.1 Learning

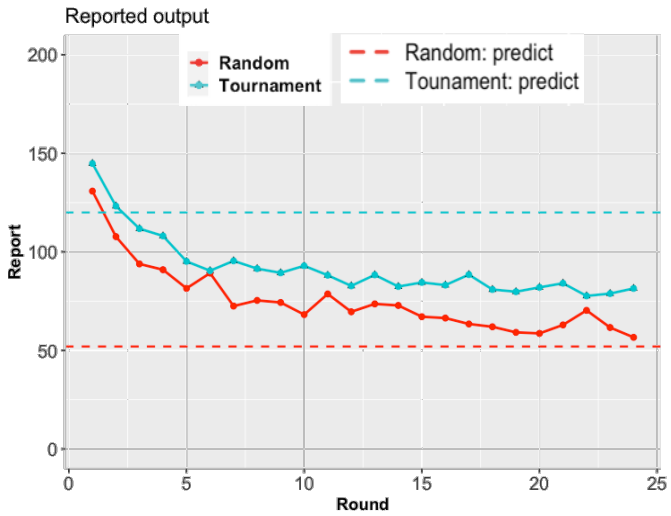
In the experiment, subjects made decisions for 24 rounds. Figure 2-1 to Figure 2-3 describe the dynamics over 24 rounds. Figure 2-1 depicts the average decisions by the audit treatments. Figure 2-2 illustrates the average decisions by the audit treatments and by the heterogeneity treatments. Figure 2-3 illustrates average decisions by the audit treatments, the heterogeneity treatments, and the different agent types. There are four plots in each figure, one for output, one for reports, one for mis reports, and one for mis reports as a percentage of total output. The dashed lines show theoretical predictions with colors corresponding to different treatments.

First, in Figure 2-1, the top two panels display results regarding the two decisions the subjects made in our experiment: output and reporting. The two lower panels used these two to compute absolute misreporting, and also misreporting as a fraction of total output. The decisions are reported by audit treatment (random versus tournament audit), as an average over homogeneous and heterogeneous agents.

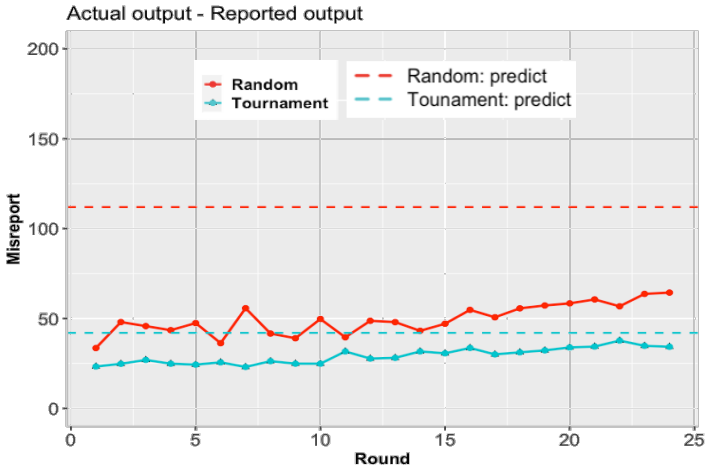
We start by discussing the output decision, which is reported in the top left figure. Notice that in the experiment, subjects can choose a level of output between 0 and 300. The dashed line shows the theoretical prediction of 161 in both audit schemes. Our results show that the output decision in the first couple of rounds tends to be higher than predicted, but then it quickly becomes lower than predicted and remains largely unchanged throughout the rest of the experiment. The average output choice in both audit schemes is approximately 24% lower than the theoretical prediction and, importantly, such under-production is consistent across audit treatments. On the other hand, as reported in the top right panel, the reporting decision gradually approaches the equilibrium prediction over subsequent rounds. As a result, both the misreporting level (bottom

left panel) and proportional misreporting (bottom right panel) show a clear trend towards the theoretical prediction as players gain experience by playing additional rounds of the experiment.





Note: theoretical predictions of the output are the same for both the random and tournament audit treatment



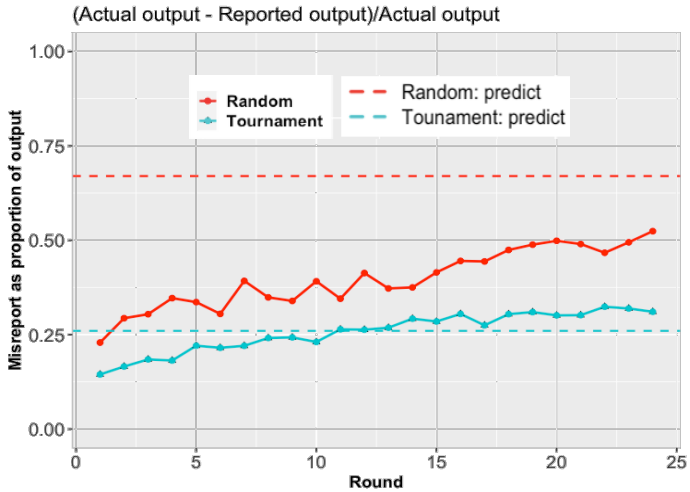
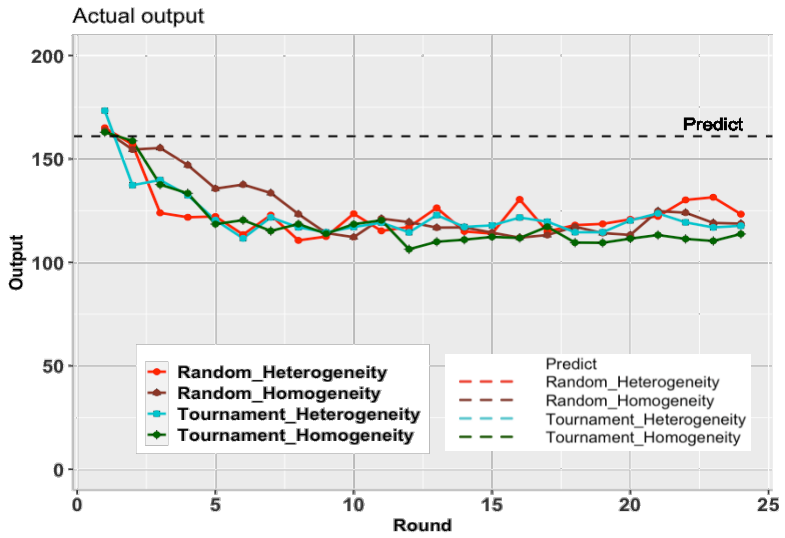


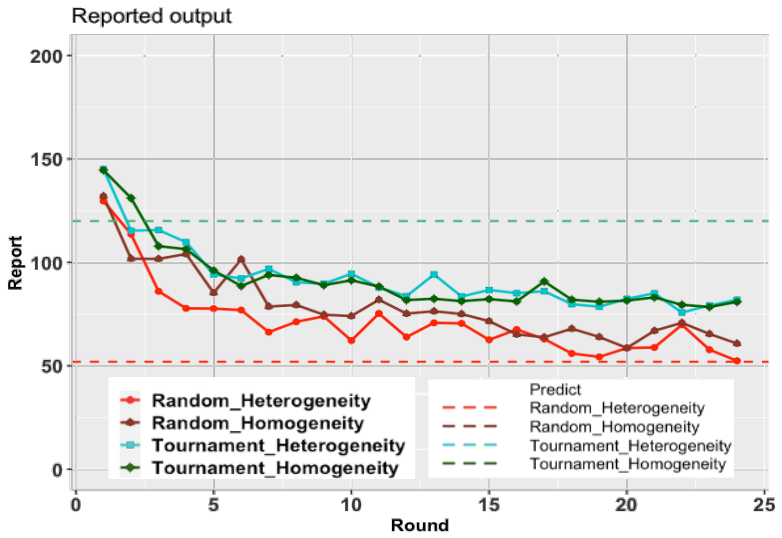
Figure 2-1 Decision by audit treatment over 24 rounds

I also find that the proportional misreporting under a tournament audit is very consistent with the theoretical prediction, while it is lower than predicted under a random audit. As a result, the experimental evidence shows that the tournament audit does not reduce misreporting as much as theory predicts it would. In absolute terms, subjects tend to under-report less than predicted. This is a finding common in regulatory compliance experiments with random audit mechanisms. For example, Alm et al (1993) and Gilpatric et al (2015) have both documented observed misreporting that is lower than predicted and pointed out that such observations may be driven by risk aversion or lying aversion. Figure 2-2 shows the average decision separately for audit treatments and heterogeneity treatments.

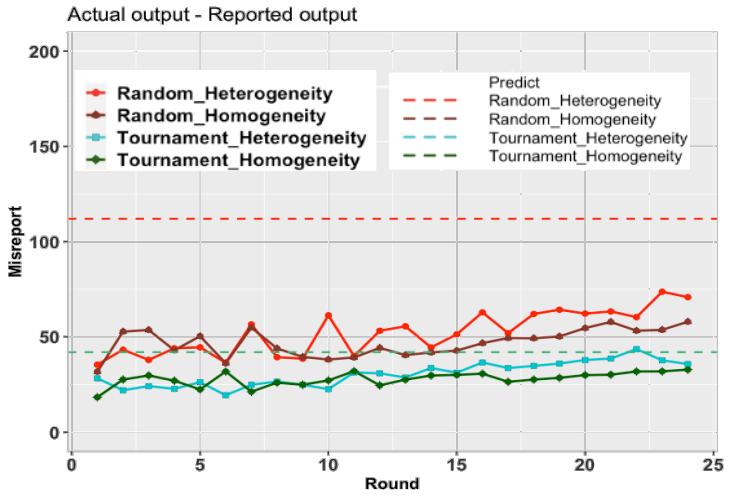
In Figure 2-3, I further disaggregate results by agent type in the heterogeneity treatment. As the top left panel indicates, output is lower than predicted, but consistent across treatments. The top right panel shows that agents report more under the tournament audit than under the random audit. Given that output is the same across audit schemes, this translates into lower misreporting in both absolute (bottom left panel) and relative terms (bottom right). As predicted by theory, the tournament audit reduces the difference in misreporting between the favorite and underdog agents, especially so for the latest rounds of the experiment. The reason for this is that, in a tournament audit, the inspection is decided based on relative suspiciousness, so the agents' strategic response is to behave similarly so as to avoid suspicion by the principal.



Note: theoretical predictions of the output are the same for both the random and tournament audit treatment



Note: theoretical predictions are the same for homogeneity and heterogeneity for each audit treatment



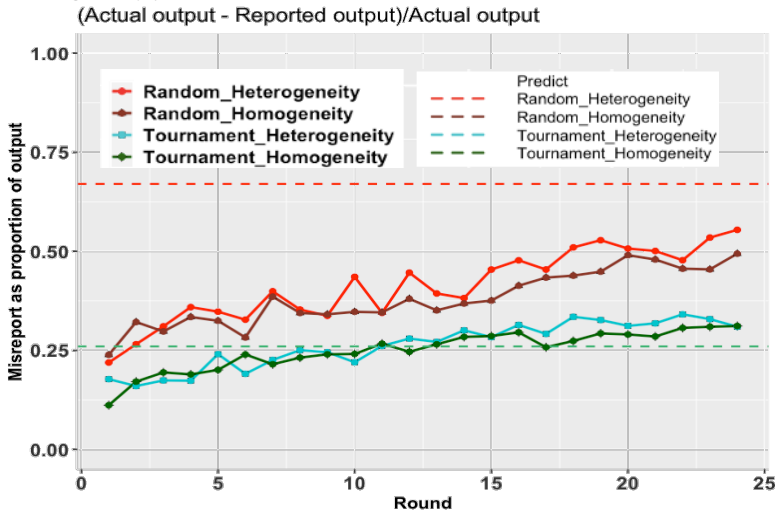
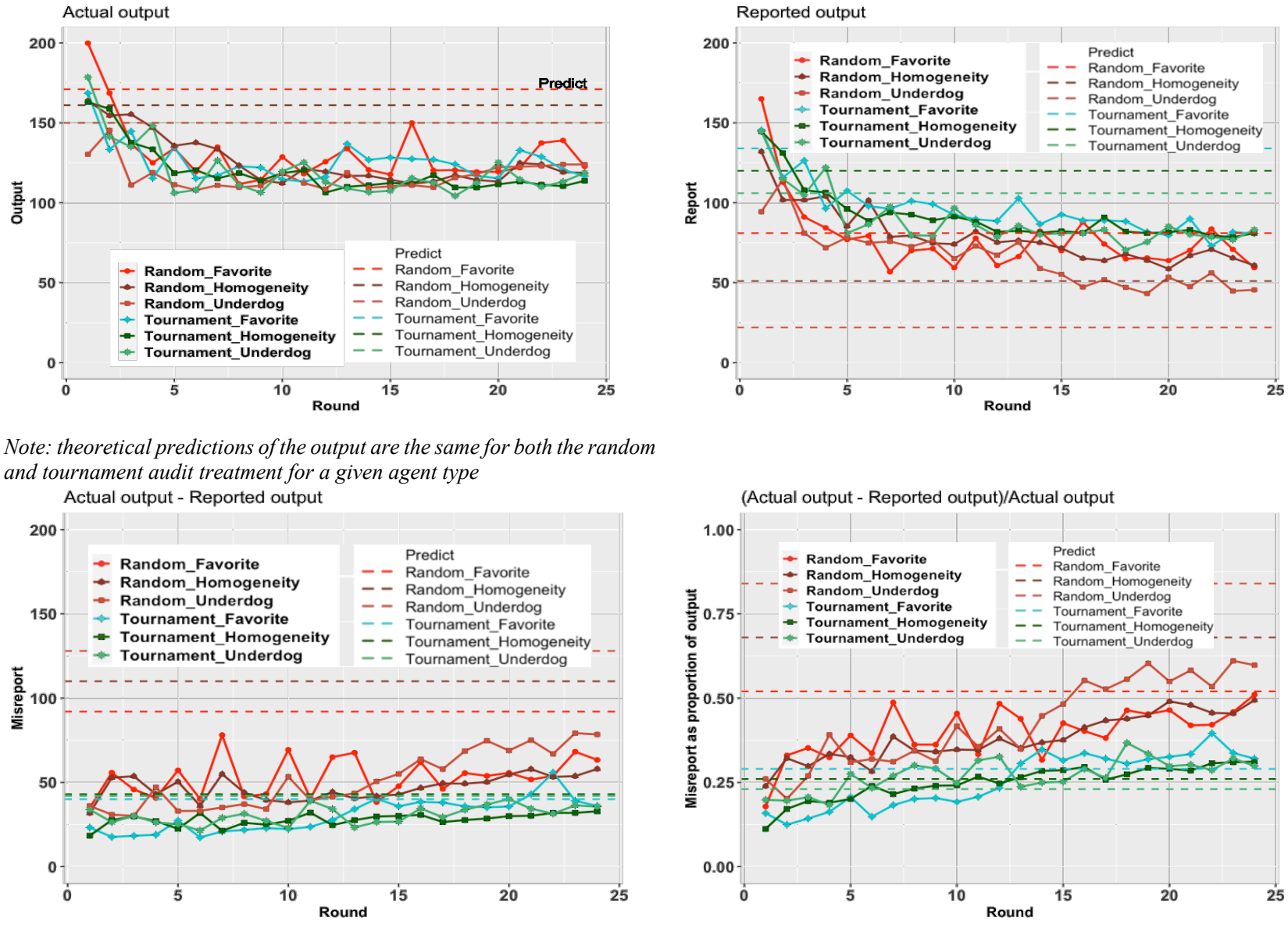


Figure 2-2 Decision by audit and heterogeneity treatment over 24 rounds



Note: theoretical predictions of the output are the same for both the random and tournament audit treatment for a given agent type

Figure 2-3 Decision by treatment and different types of subjects over 24 rounds

The results just reported are averages across subjects for a given round. These averages hide considerable variation across subjects within rounds, and also across rounds within subjects. In fact, for all the 96 subjects across 24 rounds, or 2,304 subject–round observations, only 250 subject–rounds are zero misreporting (when the subject truthfully reports all the output) while 150 are complete misreporting (when the subject dose not report any of the output). The frequency distribution of proportional misreporting shows that subjects did not use the "none or all" strategy but rather tended to find a level of misreporting that is between the zero misreporting and maximum misreporting. Figure 2-4 shows the frequency of subjects who report truthfully during a given period by the audit treatment. Under both audit rules, the majority of the subject misreport in most of the periods.

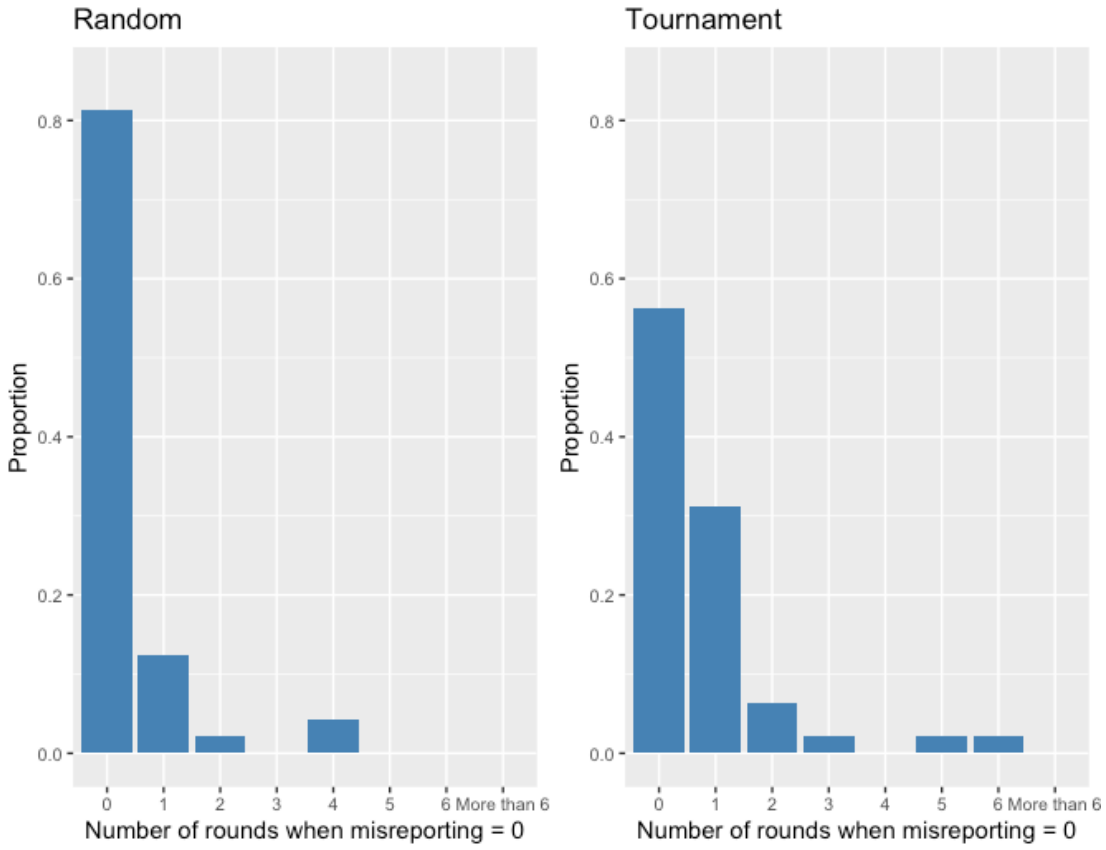


Figure 2-4 Frequency of subjects with truthful reporting rounds

4.2 Testing the main hypotheses

To test the main hypotheses, I estimate a random effect model. Table 2-5 below summarizes the number of statistically independent observations in each treatment. For the tournament audit, there

are only 2 independent observations because of the strategic interaction across subjects. In the tournament audit, the subject knows that the audit is jointly determined by everyone's behavior in each group, and the subject learns the decisions of other subjects through observing if he or she is audited, while in the random audit, each subject is an independent observation since they never learn the decisions of other subjects.

Table 2-5 Number of observation and statistically independent observation

Audit	Homogeneity/ Heterogeneity	Subject-round obs.	Interaction across the groups?	Statistically independent obs.
Random	*Homogeneity*	576	No	24 subjects
	Heterogeneity	576	No	24 subjects
Tournament	*Homogeneity*	576	Yes	2 sessions
	Heterogeneity	576	Yes	2 sessions

For the random effect model, I first use observations from all the rounds. The results are reported in Table 2-7. I them estimate the model using a subsample of observations from the last five rounds and the results are reported in Table 2-8. For both models, I examine whether the aggregate behavior conforms to the predictions of the theory in each treatment. The dependent variables in each regression are output (columns 1 and 2), reporting (columns 3 and 4), the level of misreporting (columns 5 and 6) and misreporting as a proportion of the output (columns 7 and 8).

In each regression, I include the following experimental variables: dummies for each treatment, and a *trend* variable indicating the inverse of rounds from 1 to 24. I further included the interaction terms between the two treatment indicators, and a term capturing the interaction between the inverse of the round and the *heterogeneity* treatment indicator to account for the learning process. Columns 2, 4, 6, and 8 contain indicators of favorites as well as the interaction between favorites and tournament audit treatment. I also include multiple control variables such as the subject's gender (*male=1*), major (indicators for different majors), and a discrete variable for GPA (below 2 = 1, 2 to 2.5 = 2, 2.5 to 3.5 = 3, higher than 3.5 = 4), the measure of risk preference, which is the amount of money invested in the investment task, an indicator for subjects who have experience of participating in other economics experiments, and the indicators for misreporting motives.

The first set of hypotheses is related to output. Based on the model prediction, the output choice is independent of the audit scheme. More specifically, Hypothesis 1a states that, under homogeneity, the output level chosen is the same in the random audit and tournament audit treatments. Hypothesis 1b states that, under heterogeneity, the output level chosen by agents is the same in the random and tournament audit treatments. The theoretical prediction is that output should equal 161 regardless of the audit treatment. Our estimates, reported in Tables 2-7 and 2-8, show that the average output is not significantly different across audit schemes. This result is consistent with Hypotheses 1a and 1b. Moreover, the output choice in the experiment tends to be lower than the theoretically predicted outcome. This is slightly different from Cason et al (2016) where the researchers found that output was above predicted in the random audit scheme, and below the prediction in the tournament scheme. The following are the outcomes of this discussion.

Result 1: The audit treatment does not affect the output level

The second set of results focuses on misreporting decisions. First, to examine the unconditional reduction in misreporting associated with the tournament audit. I pool misreporting decisions across homogeneity and heterogeneity treatments. This gives the average misreporting by the audit scheme. The average misreporting under the tournament audit treatment is 29.2, much lower than that under the random audit, which is 49.6. In terms of misreporting as a percentage of output, it is 25% for the tournament treatment (i.e., the average misreporting is 25% of actual output), also much lower than 38% under the random audit treatment.

Hypothesis 2a and 2b state that a tournament audit scheme reduces misreporting, regardless of agent's heterogeneity. The regression results in Table 2-7 and Table 2-8 show that the average misreporting is significantly lower in the tournament audit treatment, both in absolute terms (the misreporting is around 20 units lower in the tournament audit than that of the random audit) and as a fraction of output (misreporting is around 13% lower in the tournament audit than in the random audit). The difference in misreporting is significant at the 1% level. The magnitude of the difference is even greater for the last five rounds (Table 2-8) and remains highly significant: the average misreporting under the tournament audit is approximately 36 units (or 28% of the output) lower than that under the random audit at the 1% significance level. These results are consistent with Hypotheses 2a and 2b.

I also test if the reduction in misreporting under the tournament audit is explained by changes in the extensive margin (i.e., more subjects misreport) or intensive margin (i.e., more

misreporting by each subject). Table 2-6 shows the percentage of mis-reporters (% of total subjects who underreport) by treatment. Mis-reporters make up more than 98% of the subjects in all rounds, and 100% of all the subjects misreport in the last five rounds, indicating that the difference in average misreporting is largely due to a difference in the amount if misreporting by agents who were already underreporting instead of the proportion of subjects who underreport.

Table 2-6 Proportion of under-reporter by treatment

Audit	Homogeneity/Heterogeneity	Agent type	All rounds	Last 5 rounds
Random	*Homogeneity*	/	98.1%	100.0%
	Heterogeneity	Favorite	99.0%	100.0%
		Underdog	100.0%	100.0%
Tournament	*Homogeneity*	/	98.4%	100.0%
	Heterogeneity	Favorite	99.7%	100.0%
		Underdog	98.3%	100.0%

We summarize the above discussion with the following result.

Result 2: The tournament audit reduces misreporting compared to the random audit, regardless of the heterogeneity of agent.

We now turn to the issue of whether the experimental evidence supports or contradicts Hypothesis 2c. This hypothesis states that not only does a tournament audit reduce misreporting, but also that the magnitude of the reduction is the same under homogeneity and heterogeneity. First, I pool different types of agents together. Resulting evidence supporting this hypothesis, since the difference in the reduction of misreporting is quite small for either absolute misreporting (difference = 3) or misreporting as a proportion of output (difference = 0.02). Moreover, in the panel regression, I interact the tournament treatment indicator with the heterogeneity treatment indicator. The coefficient of this interaction term, which is interpreted as the difference in difference, is small in magnitude (-3) and is not statistically significant. This indicates that the heterogeneity treatment does not affect the reduction in misreporting from switching to a tournament audit, lending credence to hypothesis 2c. We summarize this discussion with the following result.

Results 3: The difference in misreporting between tournament and audit remains the same with and without heterogeneity (difference in difference = 0).

The final set of hypotheses focuses on the difference in misreporting between the favorites and the underdogs. To test Hypothesis 3a that the favorite always misreports less than the underdog, regardless of the audit treatment, we first look at (unconditional) results. These results show that the average misreporting of the favorite subjects is 27.9, while that of the underdog is 30.6. A similar pattern can be seen in the misreporting as a percentage of output. This is consistent with our theoretical prediction: the underdog has higher reporting costs, which induces higher misreporting. I also examine the coefficient of the *Favorite* variable in the panel regression results (Tables 2-7 and 2-8). Although this coefficient is not significant for all rounds on average (Table 2-7), it is significant at the 1% level using the last five rounds of the sample (Table 2-8). The favorite subject misreports 23 units of output less than the underdog, or 19% less as a share of output, according to results in 2-8. This result supports Hypothesis 3a.

Hypothesis 3b states that the tournament audit scheme reduces misreporting for both favorite and underdog players. Consistent with this notion, the coefficient of tournament audit treatment is significant, and the joint test for tournament audit and tournament audit interacted with the favorite indicator is also significant. On the other hand, Hypothesis 3c states that the tournament audit scheme reduces the dispersion in misreporting between the favorite and the underdog. To examine this, I consider the coefficient of the interaction term of tournament and the favorite. This coefficient has the predicted sign, but it is not statistically significant. The difference between the favorite and the underdog's misreporting is 15.9 (or 12% as a proportion of output) under a random audit scheme, but only 6.7 (or 4% as a proportion of output) under a tournament audit scheme. This suggests that strategic interaction in tournaments motivates not only lower misreporting but also smaller variance in agent misreporting, though the difference is not statistically significant (p-value = 0.28 for misreport level, 0.21 for misreport as a percentage of output).

Overall, I find that the experimental results largely support the hypothesis. I also found that being inspected in the previous round slightly increases the probability of misreporting in the current round (results not reported here). Such an observation is known as the gambler's fallacy, which happens when people mistakenly think that uncorrelated random events (for example in this case, the random audit) are correlated.

Results 4: In both random and tournament audit treatments, the favorite misreports less than the underdog. The difference during the later rounds is significant. The difference in

misreporting between favorites and underdogs is smaller in the tournament audit than in the random audit, but it is not statistically significant.

Table 2-7 Panel regression results (all rounds)

Dependent variable	Output		Report		Misreport		Misreport as % output	
Tournament = 1	-8.03	-3.91	11.53***	11.67**	-19.59***	-19.58**	-0.13***	-0.13***
	(5.96)	(4.01)	(6.52)	(6.55)	(5.01)	(5.00)	(0.04)	(0.04)
Heterogeneity= 1	-1.60	/	-8.32	/	6.72	/	0.03	/
	(6.15)		(6.70)		(5.14)		(0.05)	
Tournament Heterogeneity= 1	6.67	/	9.62	/	-3.02	/	-0.02	/
	(8.63)		(9.45)		(7.72)		(0.06)	
Favorite = 1	/	12.39***	/	9.52***	/	-3.04	/	-0.02
		(2.63)		(2.72)		(2.30)		(0.01)
Tournament Favorite= 1	/	-6.38*	/	-3.68	/	-2.84	/	-0.01
		(3.69)		(3.81)		(3.22)		(0.02)
Engineering major = 1	-6.01***	-6.13	-7.73***	-8.15	1.72	2.02	0.03***	0.041
	(2.22)	(5.87)	(-2.40)	(6.56)	(1.83)	(4.99)	(0.01)	(0.04)
Science major = 1	1.30	1.05	-3.75	-4.25	5.05**	5.31	0.03**	0.04
	(2.79)	(7.38)	(3.01)	(-8.25)	(2.30)	(6.28)	(0.01)	(0.05)
Agriculture major = 1	-37.82***	-38.01	-7.37	-7.88	-30.45***	-30.11	-0.13**	-0.12
	(8.78)	(23.27)	(9.45)	(26.01)	(7.24)	(-19.80)	(0.05)	(0.18)
Nursing major = 1	11.38*	11.88	0.18	0.46	11.19**	11.42	0.06*	0.06
	(6.21)	(16.23)	(6.69)	(18.13)	(5.13)	(-13.81)	(0.03)	(0.12)
Age	1.26**	1.30	0.99	1.01	0.26	0.29	-0.01*	-0.01
	(0.63)	(1.68)	(0.68)	(1.88)	(0.52)	(1.43)	(0.01)	(0.01)
GPA	-1.61	-1.77	1.42	1.39	-3.03	-3.17	-0.03*	-0.03
	(2.74)	(7.26)	(2.95)	(8.18)	(2.26)	(6.18)	(0.02)	(0.05)
Amount Invested	2.50***	2.49	-0.14	-0.20	2.62***	2.70*	0.01***	0.01
	(0.66)	(1.73)	(0.71)	(1.94)	(0.54)	(1.47)	(0.00)	(0.01)
1/round	65.59***	66.12***	76.29***	77.48***	-10.69***	-11.56***	-0.22***	-0.21***
	(4.76)	(4.67)	(4.21)	(4.83)	(4.15)	(4.08)	(0.03)	(0.02)
(Heterogeneity = 1) × (1/round)	-10.58	-11.95	0.75	-1.97	-11.46*	-9.60*	-0.04***	-0.04***
	(6.66)	(6.44)	(4.21)	(6.67)	(5.68)	(5.62)	(0.00)	(0.00)
Estimator	RE	RE	RE	RE	RE	RE	RE	RE
Number of Observations	2,304	2,304	2,304	2,304	2,304	2,304	2,304	2,304
R square	0.13	0.14	0.18	0.19	0.13	0.09	0.12	0.11
Adjusted R square	0.12	0.13	0.18	0.19	0.12	0.08	0.11	0.11
F	4.48***	95.16***	7.49***	259.00***	18.360***	144.400***	17.55***	290.80***

Notes: *** p<0.01, ** p<0.05, * p<0.1. Control variables also include Economics major=1, Education major=1, Male=1, results not reported due to insignificance. Baseline dummy for major: Management = 1. 1/round is the inverse of round which is 1 to 24. Unit of observations for Random Effect is subjects-round. Unit of observation for the between-estimator is subject. The results based on observations from all rounds.

Table 2-8 Panel regression results (last five rounds)

Dependent variable	**Output**		**Report**		**Misreport**		**Misreport as % output**	
Tournament = 1	-1.86	-3.49	32.59***	33.39***	-34.46***	-36.80**	-0.27***	-0.28***
	(5.35)	(4.25)	(6.98)	(6.55)	(7.91)	(5.84)	(0.05)	(0.04)
Heterogeneity= 1	0.58	/	-10.03	/	10.61	/	0.03	/
	(12.18)		(22.56)		(25.86)		(0.12)	
Tournament Heterogeneity= 1	-6.84	/	-9.14	/	2.30	/	0.49	/
	(7.73)		(9.63)		(10.92)		(0.07)	
Favorite = 1	/	-0.26	/	23.57***	/	-23.85***	/	-0.18***
		(6.09)		(2.72)		(2.30)		(0.06)
Tournament Favorite= 1	/	-9.25	/	-18.68	/	-9.44	/	-0.09
		(8.34)		(3.81)		(11.61)		(0.08)
Estimator	RE	RE	RE	RE	RE	RE	RE	RE
Number of Observations	480	480	480	480	480	480	480	480
R square	0.13	0.14	0.14	0.12	0.15	0.16	0.17	0.17
Adjusted R square	0.12	0.13	0.13	0.12	0.12	0.15	0.14	0.16
F	21***	22***	39***	53***	49***	65***	49***	58***

Notes: *** p<0.01, ** p<0.05, * p<0.1. Control variables also include Economics major=1, Education major=1, Male=1, results not reported due to insignificance. Baseline dummy for major: Management = 1. 1/round is the inverse of round which is 1 to 24. Unit of observations for Random Effect is subjects-round. Unit of observation for the between-estimator is subject. The results based on observations from the last rounds. All the control variables are the same as in Table 8. Results of these control variables are not reported here.

The above results are with regard to the efficiency improvement of the tournament audit. Based on the theoretical model, the tournament audit improves the net payoff for about 7%, both with homogeneous players and with heterogeneous players. Tables 2-9 and 2-10 show the predicted net payoff and average empirical net payoff by treatment and agent type, based on data from all rounds and the previous five rounds, respectively. The average is the highest in tournament-heterogeneity treatment for the favorites, which is consistent with the prediction. However, the average net payoff in the experiments is significantly lower than the equilibrium payoff predicted by the theory, and there is significant dispersion, as evidenced by the large standard deviations.

According to the prediction, the average net payoff in the tournament audit schemes is 8% higher than that in the random audit. Also, in the last five rounds, the empirical net payoff is much closer to the theoretical prediction. However, the average net payoffs in the tournament audit treatment are not statistically different from those in the random audit, both in all rounds and in the later rounds. This indicates that the payoff improvement effect of the tournament is not supported by the experimental results.

Table 2-9 Average net payoff: all rounds

Audit	**Homogeneity/Heterogeneity**	**Agent type**	**Predicted**	**Empirical**
Random	*Homogeneity*	/	231	67.3 (53.4)
	Heterogeneity	Favorite	265	77.6 (78.8)
		Underdog	199	80.2 (30.9)
Tournament	*Homogeneity*	/	249	74.6 (44.9)
	Heterogeneity	Favorite	278	85.2 (74.2)
		Underdog	219	49.6 (83.1)

Notes: the empirical net payoff is based on average across all rounds. Standard deviation in parathesis.

Table 2-10 Average net payoff: the last five rounds

Audit	**Homogeneity/Heterogeneity**	**Agent type**	**Predicted**	**Empirical**
Random	*Homogeneity*	/	231	107.4 (45.4)
	Heterogeneity	Favorite	265	108.8 (48.1)
		Underdog	199	84.5 (30.9)
Tournament	*Homogeneity*	/	249	109.9 (17.0)
	Heterogeneity	Favorite	278	102.4 (32.6)
		Underdog	219	85.8 (27.1)

Notes: the empirical net payoff is based on average of the last five rounds. Standard deviation in parathesis.

In terms of the net payoff distribution, Figure 2-5 shows the density of net payoff in the experiment across all rounds (top two plots) and the last five rounds (bottom two plots). For the all-round distribution, the tournament-homogeneity treatment has the least dispersed distribution. In addition, in all treatments, I see some extreme net payoffs in the left tail, which occurs when the player chooses to report no or very little output and is audited in that round. This also explains the large standard deviations reported in Table 2-10. The later rounds have higher mean and less dispersion due to the learning process, as much fewer extreme net payoffs.

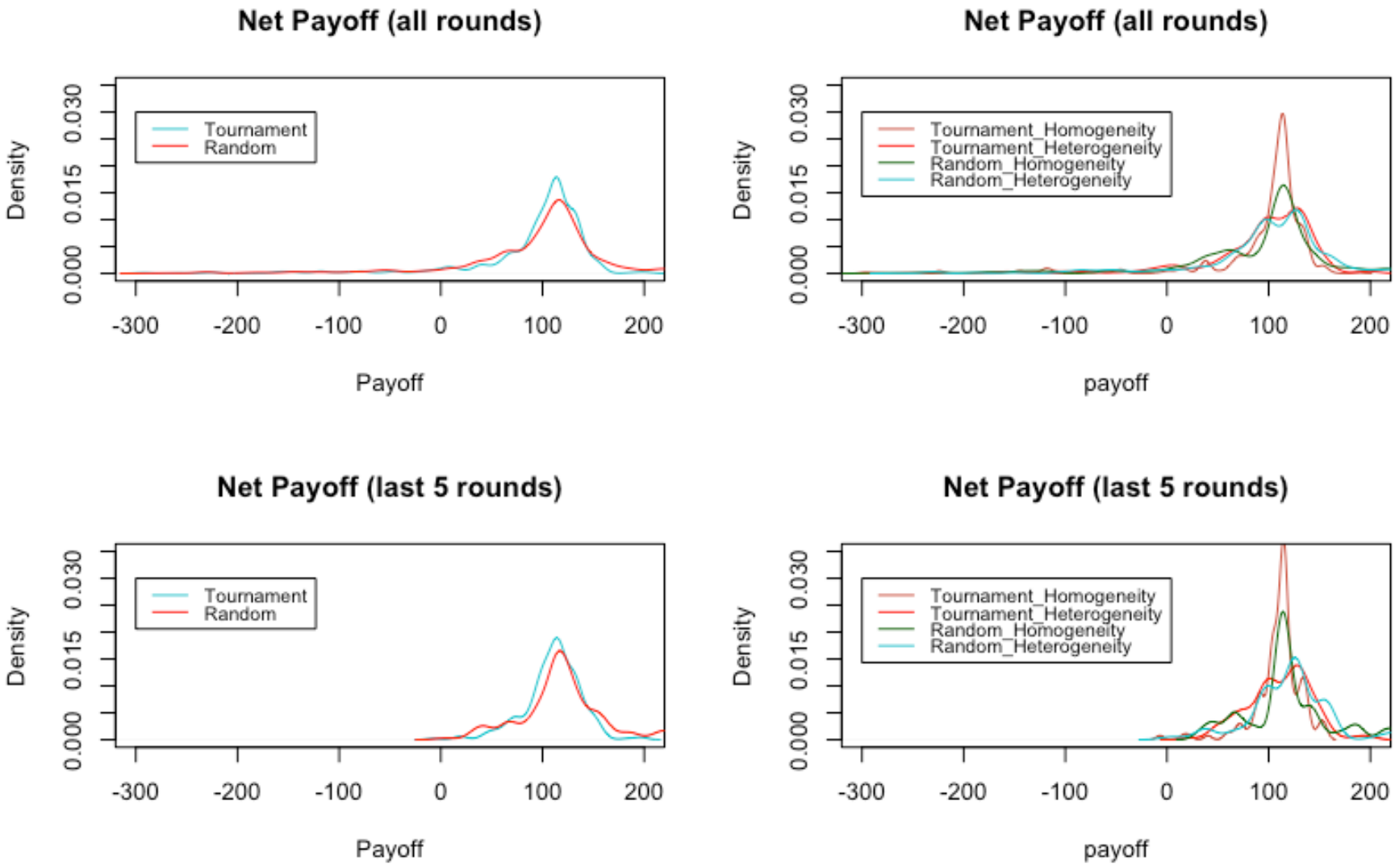


Figure 2-5 Payoff distribution by treatment

Finally, in the experiment, I observe some discrepancies and divergences from the theoretical prediction, especially for the random audit treatment. The proportional misreporting in the random audit tends to be lower than theoretical predictions. Table 2-11 shows the gap between predicted misreporting as a percentage of the output and the empirical value averaged across all the rounds. Overall, the observed proportional misreporting is lower than what is predicted by the theory, with larger under-misreporting happening for the underdog-type agent under the random-heterogeneity treatment, where the proportional misreporting is more than 40 percentage points lower than what is predicted, indicating that agents are under-misreporting considerably.

There are two possible explanations for such wide divergence. First, the experiments adopted random role switching in the blocks. The subject will be one type for the first 12 rounds and the other type for the second 12 rounds with an equal chance. Although this allows the subject to learn the game, compared to a fixed role assignment, random role switching might lose information about how subjects behave in a given role. Risk aversion might also partially explain such a divergence. As discussed in the first essay, risk averse agents lower their misreporting, especially under a random audit. With a CRRA risk aversion coefficient of 0.6, the level of misreporting and the effect of tournaments are very similar to what we observe in the lab. However, risk aversion is inconsistent with other comparative statics and the output choice in the experiment. For example, with a risk aversion coefficient of 0.6, the predicted output level is around 70, while the output in the lab is approximately 120. Also, notice that although the observed proportional misreporting is lower than predicted, the subject in the experiment still misreports a considerable proportion of their output in the random audit treatment (on average, around 40%). I further checked if the divergence differs across subjects with different characteristics, such as their demographics and their attitudes towards misreporting. The results show that none of the demographic variables are significant in affecting the size of the divergence.

The under-misreporting also leads to less payoff. Because participants tend to underproduce and underreport, this combination of strategies results in expected profits of approximately 7 experimental dollars, which is only one-third of the expected profits earned from optimal output and misreporting, which is approximately 20 experimental dollars.

Table 2-11 Misreporting as percentage of output (average of all rounds)

Audit	Homogeneity/Heterogeneity	Agent type	Predicted	Empirical	Divergence
Random	*Homogeneity*	/	67%	36%	-31%
	Heterogeneity	Favorite	52%	39%	-13%
		Underdog	84%	43%	-41%
Tournament	*Homogeneity*	/	24%	23%	-1%
	Heterogeneity	Favorite	22%	25%	3%
		Underdog	27%	26%	-1%

5 Conclusions and policy implications

This paper discusses the economics of information reporting under audit schemes based on absolute (random) and relative (tournament) misreporting. In the random audit, agents face a fixed and exogenous probability of being inspected. In the tournament audit, the relative suspicion of agents determines the chance of an audit, and the agents compete with one another by reporting strategically to avoid being audited. In this context, I am especially interested in the impact of agent heterogeneity on the patterns of misreporting and their implications for surplus and surplus distribution. From a policy making perspective, the results support the use of the tournament audit scheme both when reporting agents are identical in reporting cost and when they have different reporting costs.

Evaluating the effect of heterogeneity using experiments thus clearly distinguishes this paper from the previous tournament audits literature that assumes homogeneous agents. I consider agents vested with different costs of reporting. Such asymmetries in reporting costs are commonly observed in practices, for example for taxpayers, polluting firms, and government agencies, but have not yet been investigated formally in the context of regulatory compliance with different audit mechanisms. I model heterogeneity in a symmetric matter so that the average reporting cost parameter remains the same. This specification turns out to be more appropriate and realistic than the non-mean preserving specification of heterogeneity used in the contest literature. Intuitively, other things being equal, agents with lower reporting costs are expected to misreport less, so policies directed at reducing average reporting cost would be beneficial, but such policies may not always be available or could be too costly. It may, however, be feasible to make mean-preserving changes in reporting cost, for example, by transferring resources between players or through ability-specific sorting of players into groups.

The experimental setting follows exactly the theoretical framework from Essay 1. The findings largely support the key insights from the theoretical model on the effectiveness of audit tournaments and the impact of agent heterogeneity. The results show that the tournament audit can achieve lower individual misreporting compared to a random audit because more misreporting in tournament audit results in a higher marginal probability of audit. The misreporting behavior of favorites and underdogs in the tournament audit was similar due to two opposing forces: underdogs have a high cost to report truthfully, while favorites have a low cost. The underdog is encouraged to misreport because the high cost prevents reporting truthfully but is discouraged from misreporting by the competitive tournament mechanism. Favorites, on the other hand, are the opposite. Therefore, tournament audits can achieve similar misreporting decisions even when players are of different types. Results from all of these panel regression models with different specifications confirm those findings regarding treatment differences. Such differences are largely due to differences in the amount of underreporting rather than differences in the proportion of subjects who underreport. The absolute level of misreporting is below theoretical prediction under all treatments. The proportional misreporting as a percentage of output is quite consistent with the theoretical prediction, although subjects under the random audit misreport much less than the theory predicts, which might be explained by risk aversion.

The interaction between these two treatment dimensions can provide significant insights for both researchers and policymakers. Notice that these findings should be viewed with caution since they arise from a simple laboratory experiment, but they suggest that even with heterogeneous agents, the endogenous competitive audit rule is still able to reduce misreporting efficiently. If similar effects also exist in practice, regulators should have confidence in applying the tournament audit as an effective regulatory tool.

Because the study was undertaken in a simplified experimental setting rather than in the field, I had to omit many institutional and political details. This may lead to the question about the relevance of experimental studies for addressing real world policy issues. However, experiments are abstraction from reality and are not necessarily designed to replicate field situations by incorporating all institutional details. Rather, as pointed out in previous study, the goal of conduct experiment is to deal with general theories that economists believe should apply (Roe, 2009). If a supposedly general principle does not apply in a simplified, controlled laboratory situation, one has to question the relevance of the theory for more complex situations and for making policy

arguments (Wu, 2013). Therefore, experiments are no different from most economic studies which attempt to explain behavior using simple, stylized models that abstract from reality.

As policymakers are starting to incorporate non-random audit schemes into their compliance toolkits, it is also important to understand the underlying motivations for compliance induced by these new schemes. Findings from our research indicate that providing appropriate incentives to comply could be important in such situations. The tournament audits use a competitive and endogenous selection mechanism that relies on relative perceived underreporting amongst regulated agents. The resulting incentives lead to improved efficiency by increasing truthful reporting, thus saving valuable resources spent on audits.

Acknowledgments

This work is financially supported by the 2020 Jim and Neta Hicks Graduate Student Grants Program at Purdue University. The experimental designs and data collection methods were approved by Purdue University IRB Protocol No. 2021-424.

6 References

Alm, J., Jackson, B. R., & McKee, M. (1992). Estimating the determinants of taxpayer compliance with experimental data. National Tax Journal, 107-114.

Cason, T. N., & Gangadharan, L. (2006). An experimental study of compliance and leverage in auditing and regulatory enforcement. *Economic Inquiry, 44*(2), 352-366.

Cason, T. N., Friesen, L., & Gangadharan, L. (2016). Regulatory performance of audit tournaments and compliance observability. *European Economic Review, 85*, 288-306.

Clark, J., Friesen, L., & Muller, A. (2004). The good, the bad, and the regulator: An experimental test of two conditional audit schemes. *Economic Inquiry, 42*(1), 69-87.

Erard, B., & Feinstein, J. S. (1994). Honesty and evasion in the tax compliance game. *The RAND Journal of Economics, 1*, 1-19.

Evans, M. F., Gilpatric, S. M., & Liu, L. (2009). Regulation with direct benefits of information disclosure and imperfect monitoring. *Journal of Environmental Economics and Management, 57*(3), 284-292.

Fischbacher, U. (2007). Z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics, 10*(2), 171-178.

Friesen, L. (2003). Targeting enforcement to improve compliance with environmental regulations. *Journal of Environmental Economics and Management, 46*(1), 72-85.

Gilpatric, S., Vossler, C. A, & McKee, M. (2011). Regulatory enforcement with competitive endogenous audit mechanisms. *The RAND Journal of Economics, 42*(2), 292-312.

Gilpatric, S., Vossler, C., & Liu, L. (2015). Using competition to stimulate regulatory compliance: a tournament-based dynamic targeting mechanism. *Journal of Economic Behavior & Organization, 119*, 182-196.

Greiner, B. (2004). An online recruitment system for economic experiments. *Forschung und wissenschaftliches Rechnen, 63*, 79-93.

Harrington, W. (1988). Enforcement leverage when penalties are restricted. *Journal of Public Economics, 37*(1), 29-53.

Macho-Stadler, I., & Perez-Castrillo, D. (2006). Optimal enforcement policy and firms' emissions and compliance with environmental taxes. *Journal of Environmental Economics and Management, 51*(1), 110-131.

Mookherjee, D., & Png, I. (1989). Optimal auditing, insurance, and redistribution. *The Quarterly Journal of Economics, 104*(2), 399-415.

Oestreich, A. M. (2015). Firms' emissions and self-reporting under competitive audit mechanisms. *Environmental and resource economics, 62*(4), 949-978.

Roe, B. E. (2009). Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics, 91*(5), 1266-1271.

Schotter, A., & Weigelt, K. (1992). Asymmetric tournaments, equal opportunity laws, and affirmative action: some experimental results. *The Quarterly Journal of Economics, 107*(2), 511-539.

Stowe, C. J., & Gilpatric, S. M. (2010). Cheating and enforcement in asymmetric rank-order tournaments. *Southern Economic Journal, 77*(1), 1-14.

Tullock, G. (1980). Efficient rent seeking. In *In Efficient Rent-Seeking* (pp. 3-16). Boston, MA: Springer.

7 Appendix

Appendix A. Experiment instructions

Experiment Instructions *(tournament-heterogeneity treatment)*

Introduction

Thank you for participating in this experiment on individual decision making. The amount of money you earn depends partly on the decisions you make thus you should read the instructions carefully. The money you earn will be paid privately to you, in cash, at the end of the experiment. Please put away your cell phones. Please do not communicate with other participants during the experiment. If you have a question as I read through the instructions or any time during the experiment, please raise your hand and I will come by to answer it. At the end of these instructions, you will take a quiz and earn 1 U.S. dollar for each correct answer you provide. You may refer to the instruction anytime.

The experiment is divided into 24 rounds. You will be paid based on the sum of your earnings from 8 randomly chosen rounds. Your earnings in the experiment are in experimental dollars, which will be exchanged at a rate of 80 experimental dollars = 1 U.S. dollar.

Each round, you will be in a group consisting of 4 members, you and 3 others who are also sitting in this room. Your will be randomly matched to 3 members in each round. They may or may not be the same people you have interacted with in previous rounds. You will make decisions privately without consulting other members.

Overview

Each round, you will decide how much output to produce and then decide how much to report. The higher you produce, the higher the revenue and the higher the production cost. The higher you report your output, the higher the reporting cost.

The difference between your actual output and reported output affects the chance that you will be inspected. 2 out of the 4 members in the group will be inspected. If you are inspected and found to have reported less than your actual output, you will face additional costs. Your earnings in one round depend only on your decision- and the decisions of others- in that particular round.

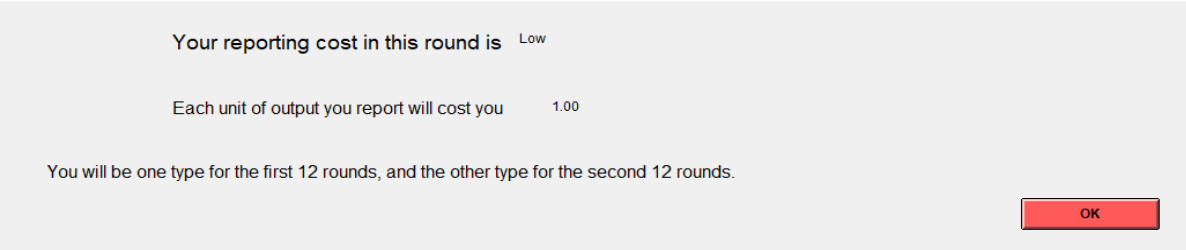
Your earnings in each round = Revenue from output – Cost from output – Reporting cost –

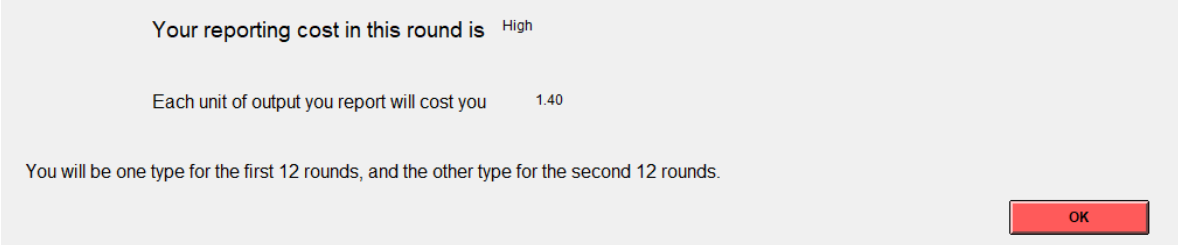
Additional costs if you are inspected and your reported output is less than your actual output.

Your reporting cost can be High or Low. There will be 2 High-cost types and 2 Low-cost types in your group. You will be one type for the first 12 rounds, and the other type for the second 12 rounds.

For example, you can be a High-cost type for the first 12 rounds and a Low-cost type for the second 12 rounds. Or you can be a Low-cost type for the first 12 rounds and a High cost type for the second 12 rounds.

This information will be shown on your screen like the one below.

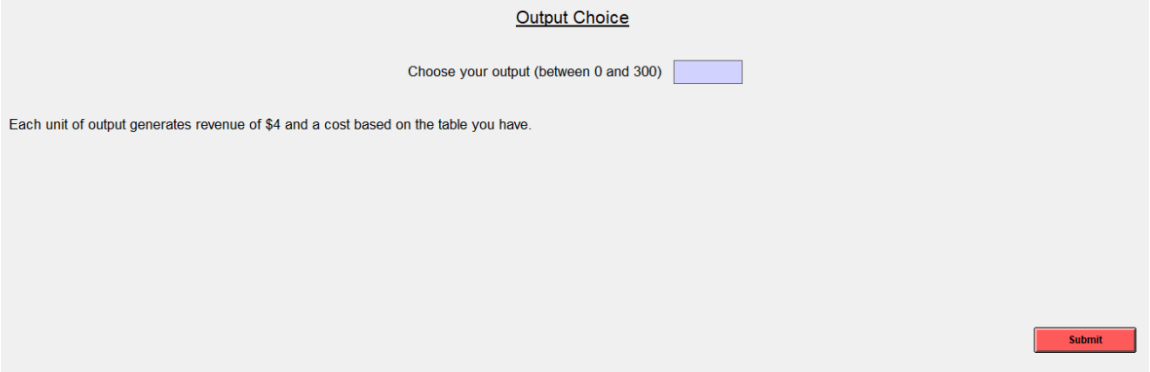




Your decisions

In each round you will make two decisions.

Your first decision is to choose an amount of output to produce. This amount must be between 0 and 300. Your decision will be entered on a screen like the one below.



Each unit of output generates a revenue of 4 experimental dollar for you. The higher the output you choose, the higher will be your revenue and the cost. Figure 1 shows the revenue and cost at different levels of output.

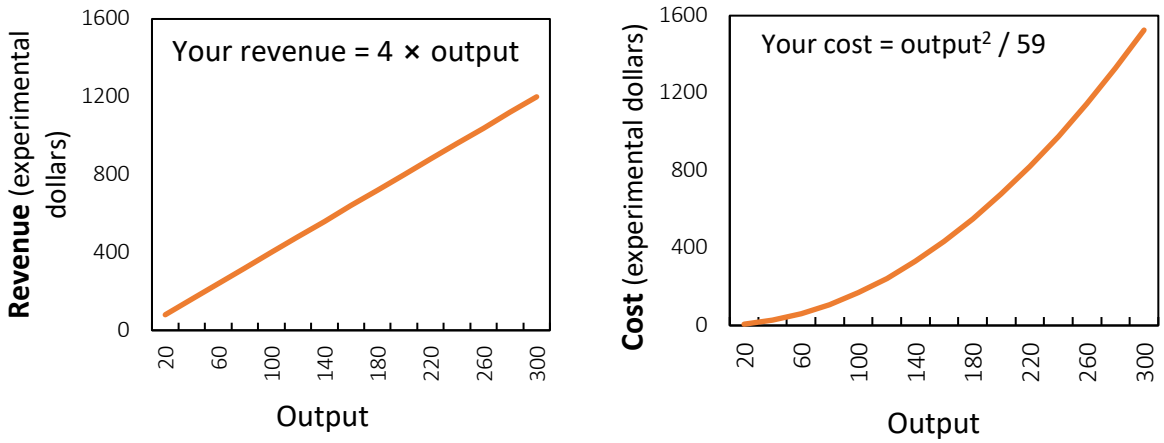
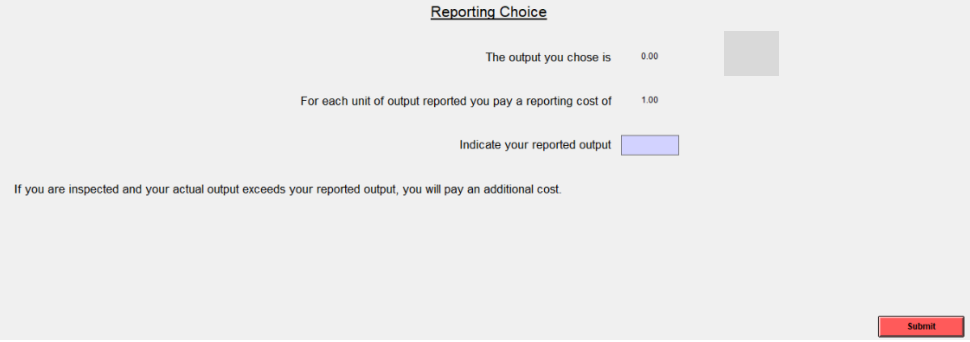


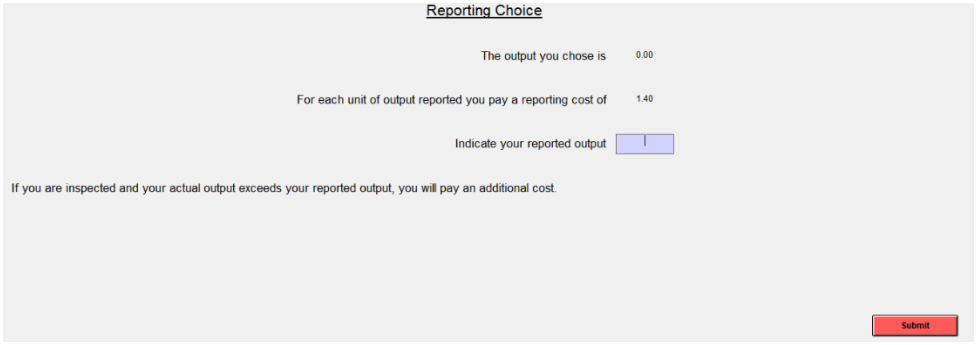
Figure 1. Your revenue and cost at different levels of output

Your second decision is to choose how much output to report. For each unit of reported output, you pay a reporting cost. Your decision will be entered on a screen like the one below. You can choose to report any amount you like, such as your actual output or an amount less or more.

- If you are a Low-cost type, you pay **1** experiment dollar for each unit of output reported.



- If you are a High-cost type, you pay **1.4** experiment dollar for each unit of output



reported.

After you have submitted your report, three things can happen:

1. You are not inspected. In this case you do not pay any additional cost.
2. You are inspected and your actual output is less than or equal to your reported output. In this case you will not pay the additional cost.
3. You are inspected and your actual output is greater than your reported output. In this case you will pay the additional cost.

Determining who is Inspected

Once you have submitted your report, the computer will estimate your output. You will NOT know the estimated output. This estimate is equal to your actual output plus a random amount. The random amount has an equal chance of being any integer between and include -90 and 90. The random amount is equal to 0 on average, which means on average estimated output is equal to your actual output.

The computer will rank all 4 members based on the estimated output minus their report. The 2 members with the greatest gap between the estimated output and their report will be inspected. The other 2 members will not be inspected in that round. Table 2 shows an example.

Table 2. Illustrative example of inspection in one round

	Actual Output	Reporting	Output estimated by the computer	Gap estimated by the computer	Inspection
	A	**R**	**A +** ***random amount***	**A + random amount - R**	
Member 1	100	95	100 - *11* = 89	89 – 95 = -6	No
Member 2	200	200	200 + *63* = 263	263 – 200 = **63**	**Yes**
Member 3	152	111	152 – *39* = 113	113 – 111 = 2	No
Member 4	80	70	80 + *25* = 105	105 - 70 = **35**	**Yes**

What happens if you get inspected?

If you get inspected, the inspection reveals your actual output. If your actual output is greater than your reported output, you pay the additional cost. The larger the gap between your output and your reported output, the higher the additional cost.

- If you are a Low-cost type, the additional cost if you get inspected is:
  $1 \times$ (Actual output – Reported output) + [(Actual output – Reported output)^2 /94]
- If you are a High-cost type, the additional cost if you get inspected is:
  $1.4 \times$ (Actual output – Reported output) + [(Actual output- Reported output)^2 /94]

For example, if your actual output is 137 and you reported 114, the gap will be 137-114 = 23

- If you are a Low-cost type, the additional cost if you get inspected is:
  $1 \times 23 + 23^2 /94 = 28.6$ experimental dollars
- If you are a High-cost type, the additional cost if you get inspected is:
  $1.4 \times 23 + 23^2 /94 = 37.8$ experimental dollars

Results

After all the members of your group have made their decisions, you will see the results screen. It displays whether you were inspected and your earnings in that round.

After the experiment

After the experiment, you will participate in an investment task. You will receive 5 U.S

dollars and have the option to invest. You will see more detail of the investment on your computer screen during the task.

Your final payment

Your final payment = *7 U.S dollars Show up fee + Quiz earnings + earnings from 8 randomly chosen rounds of the experiment + investment earnings*

If the sum of your earnings from those 8 rounds is negative, it will be counted as 0 and you will still be paid your show-up fee, quiz earnings, and your investment earnings.

Appendix B. Summary statistics

Table 2-12 Summary statistics of subject's characteristics (N = 96)

Variable	***Value***	***Random***		***Tournament***	
		Homogeneity	Heterogeneity	Homogeneity	Heterogeneity
Gender	Male	11	13	12	12
	Female	13	11	12	12
Age	/	21.1	21.4	21.3	21.4
Major	Management/Business	6	7	9	2
	Economics	1	0	2	3
	Humanities	0	0	0	0
	Liberal Arts	0	0	0	0
	Education	0	0	1	0
	Engineering	7	7	8	8
	Science	6	5	1	3
	Social Sciences	0	1	0	0
	Agriculture	0	1	0	0
	Pharmacy	1	2	0	1
	Nursing	0	0	2	0
	Other major	3	1	1	7
GPA	3.5-4	13	17	15	10
	3-3.49	11	4	8	9
	2.5-2.99	0	2	1	5
	2-2.49	0	0	0	0
	< 2	0	0	0	0
	Not applicable	0	1	0	0
Year in college	First year	0	0	2	0
	Second year	3	7	5	3
	Third year	10	7	3	6
	Fourth year	10	6	12	13
	Graduate	1	4	2	2
Number of experiments before	None	0	3	3	2
	1 to 2	6	6	5	2
	3 to 5	6	2	7	10
	More than 5.	12	13	9	10

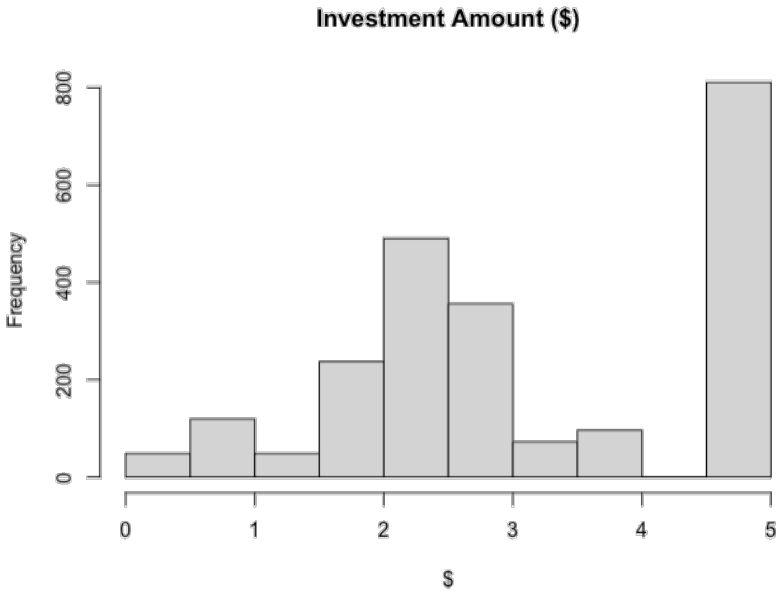


Figure 2-6 Amount invested in the risk preference elicitation task

After the experiment, the subjects were also asked a set of questions about their misreporting motives. The questionnaire asks to what extent is penalties/compliance/violation an important motivator for the choices they made during the experiment. The following pie chart summaries how the 96 subjects think about lying and obedience behavior. 77% of the subjects somewhat agree or strongly agree that it is important for them to obey all laws. A smaller proportion (49%) somewhat agree or strongly agree that it is ok to disobey laws that they do not agree with. Interestingly, more than half subjects somewhat agree or strongly agree that they lie when either the risk of being caught is low (54%), or when the consequence of being caught is low (56%)

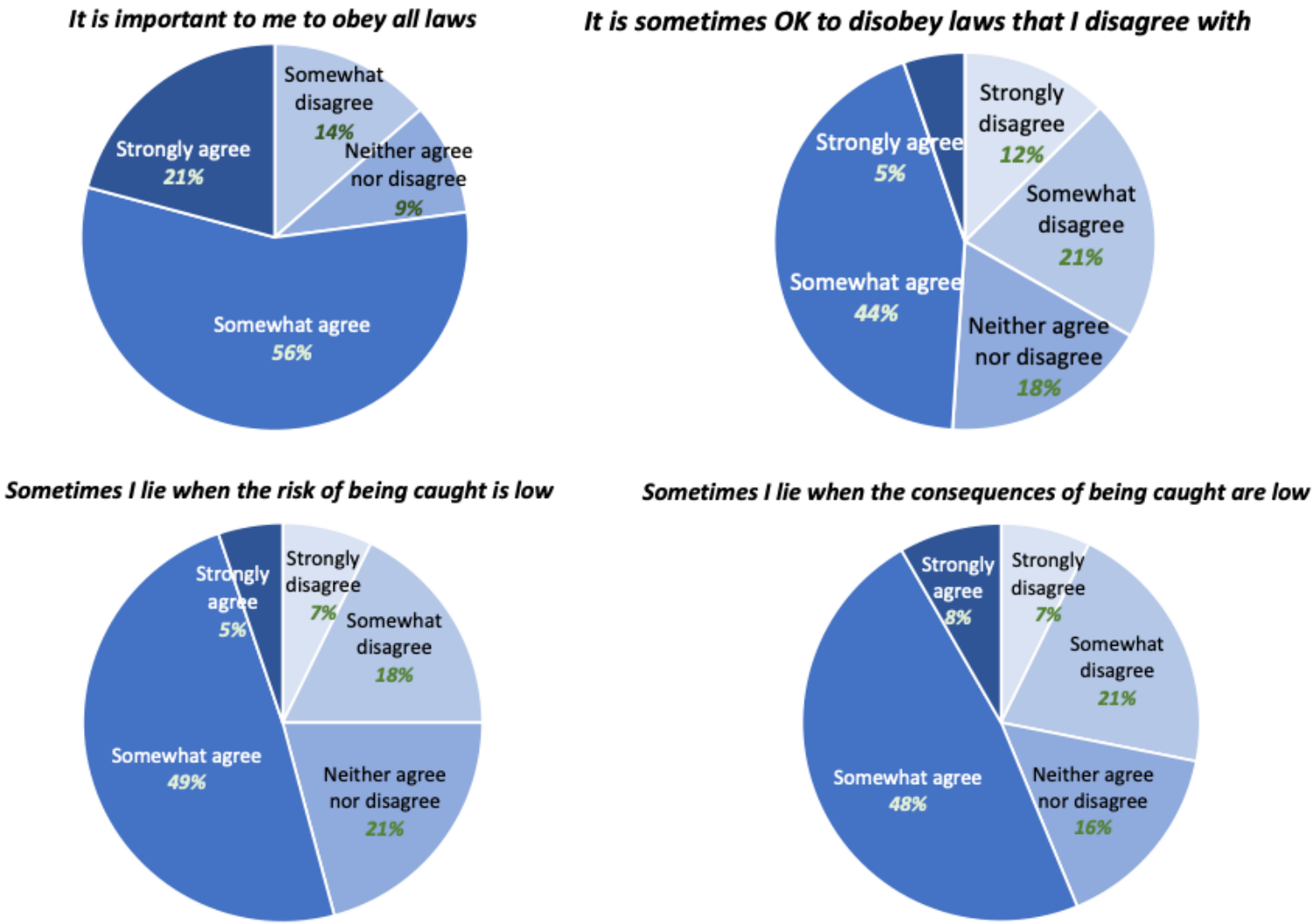


Figure 2-7 Misreporting motives

ESSAY 3. ENVIRONMENTAL CENTRALIZATION AND LOCAL AIR POLLUTION DATA MANIPULATION

Abstract

Central authorities often delegate environmental management to local authorities. Attaining a certain environmental target can be costly for local authorities and performance may be hard to monitor for supervisors or central authorities. This combination has the potential to cause local governments to reduce effort and underreport performance. Yet, there is a dearth of empirical studies measuring the extent to which enhanced monitoring leads to a reduction in misreporting. We examine this issue by studying a recent environmental reform in China. The reform gave the provincial authorities direct access to pollution monitoring stations, thereby making it harder for local authorities to misreport them. We exploit the fact that the reform was only implemented in a subset of cities to identify the impact of direct monitoring on misreporting. We do so by using DID estimators with a spatial lagged air pollution term to account for inter-region dispersion. We find that this reform significantly reduced *hard misreporting*–a reduction in measured pollution attained by interfering with the monitoring stations. On average, we find that the reform reduced hard misreporting by 2%. On the other hand, we do not find evidence of a reduction in *soft misreporting*–a reduction in measured pollution attained by altering or modifying the pollution data after pollution was measured by the monitoring stations. These results suggest that enhanced monitoring can in fact reduce misreporting.

1 Introduction

Upper-level authorities often delegate management of key resources to local authorities. Natural resources and, more generally, the environment, are typical examples. Delegation presents a tradeoff: local authorities typically have better knowledge of the resources, but also fail to internalize the externalities imposed on other districts. In theory, if attaining a certain environmental goal is costly for the local authority (because of the financial cost involved or opportunity cost, for instance in the form of reduced economic growth) and/or has limited benefits (because of free riding by other local authorities), then the central authority can reward the local authority based on environmental performance (Hölmstrom, 1979). Unfortunately, the

performance of the local authorities is often not directly observable. When attaining an environmental goal is costly and performance is costly to monitor, the local authority may be tempted to curb effort and misreport its performance. In this paper, we examine the extent to which increased monitoring reduces misreporting (e.g., Mitnick, 1980; Wilson, 1989; Williamson, 1996; Aghion and Tirole, 1997; Blonz, 2019).

A recent environmental reform in China, called the environmental vertical reform, provides an appealing empirical setting to study the effect of increased monitoring on reducing agents misreporting. In recent years, reducing air pollution has become an important focus of the Chinese central government and such a shift in focus has translated into stricter environmental targets. This incentivizes city governments to skew air pollution statistics to meet those targets, making air pollution data vulnerable to manipulation. In fact, many local governments have been recently caught manipulating pollution information. Previous studies and those with detailed knowledge of the institutions of environmental management in China have identified two pathways for manipulating pollution information. The first, henceforth called *hard misreporting*, consists of artificially reducing pollution around the monitoring stations. This involves covering air filters or spraying water around the monitoring stations, spraying chemical solutions on the station to neutralize acid pollution particles, and planting trees around the station, among others. The second, henceforth called *soft misreporting*, consists of modifying or falsifying the pollution data reported by the monitoring stations. This involves manipulating data so that reported averages are below the target set by the central authority.

Misreporting, of either type, can substantially hinder the central government's ability to evaluate local environmental quality and enforce environmental targets. As a result, the Chinese government has instituted an environmental vertical reform. According to the official guideline document, which was released in 2016, the power to collect environmental information is no longer *de jure* designated to the city government and allows the provincial and central governments to access and manage data directly (Addaney, 2017; Ma, 2017; Ran, 2018; Zhou, 2020; Yang, 2020). A key issue that assists in econometric identification is that the reform was only introduced in some regions in China, delivering a control group against which treated cities can be compared to. This allows us to quantify the effect of the reform on misreporting.

Our strategy to quantify the effect of the reform on misreporting consists of two steps. In the first step, we compute misreporting, and in the second step, we examine the effect of the reform

on measured misreporting. A key challenge in the first step is that different types of misreporting may introduce different distortions in the reported pollution data. Around a specific monitoring station, pollution can vary widely in short periods of time (hours, days, and weeks), creating a distribution (a probability density function) of pollution levels along a domain of possible values. Under soft misreporting, local authorities may artificially lower pollution measurements that are just above but close to the cutoff so that they fall below, thereby creating a discrete jump in the distribution around the cutoff.[6] To compute soft misreporting, we use a regression discontinuity design. Under hard misreporting, on the other hand, local authorities take actions that lower pollution levels around monitoring stations or limit the stations' ability to correctly measure pollution, thereby creating a shift in the entire distribution. We compute hard misreporting as the deviation between reported pollution and AOD data from NASA's satellite, which we take as a (proxy of) an unbiased benchmark.

The reform was only implemented in a subset of Chinese cities, but cities are not likely to have been randomly selected into the reform. To address this selection problem, we use the difference-in-difference (DID) and DID-matching strategies to maximize the overlap between the treated and the control groups. Pre-treatment characteristics of matched observations strongly support the parallel trend assumption. Moreover, we do not find significant evidence of spatial spillovers from treatment, suggesting no violation of the Stable Unit Treatment Value Assumption (SUTVA). All this lends credence to our DID approach and our results.

Our findings indicate that the reform had a significant impact on reducing hard misreporting. Misreporting caused by interfering with the station was reduced by about 2 percentage points after the reform. Through the quantile regression analysis, we uncover evidence that is mostly driven by a reduction in misreporting when pollution is severe. In contrast, the reform does not seem to curb data manipulation around the environmental target (cutoff) set by the central government. This does not seem problematic, however, given that such manipulation was limited before the reform was implemented. By examining the evolution of this type of manipulation, we also find that it had been substantially curtailed before the implementation of the reform, largely through the installation of upgraded automatic reading equipment since 2013.

[6] They are less likely to lower observations that are significantly above the cutoff, as these distortions may be more suspicious and easier to detect.

Although the empirical strategy relies on this particular reform in China, the findings from this paper are broadly related to the striking issue of government misreporting around the world. In fact, government data fraud happens in both developing and developed country governments and democratic countries are not immune from the temptation to misreport. Notable scandals where misreporting was detected include Greece, Bulgaria, Hungary (Carassava, 2010), Spain (European Council, 2015), and Austria (European Council, 2018) which underreported deficit or debt to GDP ratios; Puerto Rico underreported a death toll (BBC, 2018); Argentina underreported inflation rates (LeBeau, 2018); the Japanese department of labor overreported local wages (Tetsushi, 2019); Rwanda underreported poverty (A straightforward case of fake statistics, 2019); India overreported local GDP (Kumar, 2019). Martinez (2018) discovers that, after controlling for differences in economic structures, human capital factors, and so on, autocratic countries have higher GDP growth for the same amount of economic activity quantified by night light, which could lead to different translations between democratic and autocratic countries. The results suggest that the official GDP in autocratic countries is likely to have been exaggerated. Amid the COVID, many official statistics related to COVID cases have also been questioned.

The policy implications of this paper could provide important insights into the design of a misreporting reduction strategy. There has been an increasing anti-misreporting effort because government misreporting can lead to severe consequences. On the one hand, truthful information revelation is the key to accountability, and it is important to maintain trust in, and the credibility of, official statistics. Misreporting by the government has a number of negative consequences. Some recognize government data misrepresentation as to the combination of "service delivery failure" and "accountability failure", hindering the quality of public services. In environmental economics, manipulation in principle-agent settings constitutes a major challenge to the effective implementation and evaluation of pollution control policies (Ghanem, Shen, & Zhang, 2020).

This paper is most closely related to previous papers that also evaluate the effect of anti-misreporting policies or initiatives. There have been different efforts around the world to incentivize statistical integrity and the transparency of official data. These efforts either rely on better information collection technology (Greenestone et al 2020) or better public access to

government data. (Worthy, 2015; Williams, 2015; Glennerster and Shin, 2008)[7]. Greenstone et al (2020) studies the impact of better data collection technology on reducing data misreporting. The authors examine the installation of automatic air pollution monitoring. Monitoring automation aims to provide reliable measurements of pollution to identify local officials' success in achieving their targets and where a more stringent policy is necessary. They find that reported air pollution concentrations increased significantly immediately post–automation, suggesting less pollution underreporting. Worthy (2015) evaluates the democratic impact of the UK coalition government's Transparency Agenda, an open data reform aimed at improving government transparency through publishing and sharing government data. Under this reform, almost all local government spending records are shared through online information systems. The author conducted a survey of various data users, and the results show that, despite the positive impact of Open Data Reform, the complexity and fragmentation of this data limits its success in a broader sense and may even cause discontent or resistance among public bodies. McGee & Gaventa (2011) review citizen-led initiatives that aim at improving public access to official information, government transparency, and accountability. They focus on movements that are social and bottom-up, rather than those that are bureaucratic. Glennerster and Shin (2008) reveal that countries that adopted reforms that enhanced fiscal transparency experienced a structural downward shift in their credit spreads, and increased transparency is particularly beneficial for countries with smaller and less liquid debt markets. There has been an increase in the number of studies on the impact of this reform within this strand of research. Previous studies have been mainly qualitative and narrative (Addaney, 2017; Ma, 2017; Ran, 2018; Zhou, 2020; Yang, 2020). Yet none of the studies have assessed the effect of centralization reform on local misreporting. Thus, this paper extends these studies by focusing on the impact of centralization, a channel that has yet been examined empirically.

This paper also extends the empirical evidence on government misreporting. Related studies include Martinez (2018), Edmond & Lu (2018), Kalgin (2016), Ghanem & Zhang (2014). These studies have focused on how to detect a particular type of manipulation. In this paper, I use these newly developed methods to quantify manipulation and then assess the reform's impact on reducing misreporting. In addition, past studies have mainly looked at one particular type of

[7] For example, UK implemented ODA (Open Data Access) reform in 2013 to improve government information accountability (Worthy, 2015), and China adopted better monitoring technology that enables automatic readings (Greenstone, 2020).

misreporting, while in this paper, I combine several different methods for identifying different types of misreporting, including misreporting that may nor may not result in discontinuity and misreporting that happens at the tail of the air pollution distribution.

Given the pervasiveness of governmental misreporting in developing countries, the high cost of verifying local information, and the severe consequences of government misreporting, this paper offers some empirical guidance regarding designing an institutional structure that better incentivize truthful revelation. Evaluating the effect of such a reform on government misreporting could inform the design of a more institutionally realistic authority structure that has a higher potential for being successful in reducing political opportunism. In the next three sections, I introduce the institutional background. In particular, in section 2, I introduce the power structure before the environmental centralization reform, in section 3, I introduce two types of the misreporting as the consequences of such a system, and after that in section 4, I introduce the reform and how it changes the institutional structure.

2 Environmental administration structure in China

In this section, I describe the structure of the Chinese environmental bureaucracy before the reform and how it opened door to data misreporting. Before the reform, the environmental administration in China was highly decentralized. This is part of China's several major institutional transformations since the year 1978, the beginning of a series of economic and government reforms. As a major part of the overall transformative process, government decentralization transfers most of the decision-making power to the local government, which is accountable for local affairs and has considerable discretion over its fiscal revenue and expenditure.

Along with the decentralization process, China also adopts target-based yardstick competition to motivate local leaders, and leaders compete for limited promotion opportunities through outperforming other leaders (Li & Zhou, 2005). Decentralization and yardstick competition, these two important institutional elements, together strongly incentivize local officials to stimulate local development, including economic growth, social stability, and, more recently, environmental performance. (Cai & Treisman, 2006).

The local government in China is comprised of a territorial government (for example, the Beijing City Government) and functional units (for example, Beijing Environmental Protection Department). A decentralized system is characterized by the considerable power that the territorial

government has over the functional unit at the same level. As an example, city environmental protection bureaus (the functional units) are administered by the city government (the territorial government). In fact, any functional units are not allowed to issue binding orders to the territorial government of the same jurisdiction. Thus the territorial government is considered at a higher administrative rank than the functional units (Lieberthal & Oksenberg, 1990). The environmental bureaucracy in China before the 2016 vertical reform was also characterized by such a power structure.

Before the reform, China had established a dedicated environmental bureaucracy extending from the central environmental department, or MEP (Ministry of Environmental Protection, currently renamed as the Ministry of Economy and Environment, MEE) down through provinces, cities, counties, and townships. The municipal level environmental agency is called the Environmental Protection Bureau, or EPB. The administrative structure is called dual leadership, because the city's EPB is subject to both the leadership of its functional superiors, the provincial EPB, and to its jurisdictional superiors, the city government (Figure 5, top panel). The city government has the authority to determine the officials of the city EPB and approve its budget. The provincial EPB offers policy guidance and technical support. Consequently, the city government holds the dominant leadership power while the provincial EPB only plays an auxiliary and supporting role.

This decentralized system has significant advantages in that it allows local governments greater flexibility in developing environmental policies (Eaton & Kostka, 2014). This is especially important for countries like China, because decentralization allows policy heterogeneity across localities, supporting local governments in making specific environmental decisions that fit local conditions (Oates 2005). It is also efficient fiscally by making the funneling of capital more effective. Under the tax division system, each locality would pay tax to the central government. Part of this tax was returned to local governments to finance the execution of policies. From there, te money set aside for environmental protection was then distributed to the separate agencies. There was a lot of bureaucracy to go through since each agency would individually have to report to its parallel local government branch when applying for funds, and one of the intentions of decentralized management is to make the funneling of capital more direct (Ma Y. , 2017). From the perspective of the central government, decentralization allows it to distance itself from blame-generating situations (Weaver, 1986). Although such a structure makes it more difficult for the

central government to claim credit, it also helps to minimize blames leaving the central government in a safer place against public dissatisfaction (Ran, 2017).

However, this decentralized environmental management system also results in considerable negative effects. One immediate drawback is creating negative externalities. Local protectionism arises as environmental policies are made within each jurisdiction and interregional pollution becomes a pervasive issue. Moreover, since local governments have discretionary power, if they also face pressure to boost the local economy, environmental protection will never rank very high on their list of concerns. Most environmental objectives are too complex, long term, and essentially conflict with economic objectives to be effective. When the central government, or top leaders, express their concern about a certain issue, local officials will turn their attention to the issue and treat it carefully. When the attention at the central level shifts elsewhere, local compliance quickly falls (Lieberthal & Oksenberg, 1990). This is a long-lasting dilemma for China: dramatic economic growth but overwhelming problems with the environment (See Jia, 2013, for a detailed discussion).

Decentralization also makes it more challenging for the upper-level government to monitor local actions, and a policy implementation gap has been observed for many central initiatives. In terms of air pollution, the central government has issued several comprehensive guidelines on air pollution control since 2003, but the city government appears to only selectively implement those short-term policies. Compared to fast measures such as vehicle control, long-term approaches that have been strongly promoted by the central government, such as economic restructuring, energy upgrading, etc., are much less adopted (Eaton & Kostka, 2014).

Finally, one major disadvantage of the decentralized system is that it opens the door to opportunism. Poor environmental performance may adversely influence many aspects, including city leaders' careers, China launched the War on Air Pollution in 2012. Since then, the Chinese central government has put more emphasis on local leaders' environmental performance. For example, in the 12th and 13th Five-Year Plans, which offer general guidelines that cover the period 2010–2015 and 2016–2020, respectively, both plans set air pollution reduction targets for different areas and urge local leaders to improve air quality in their jurisdiction. In addition, in 2014, the National Air Pollution Prevention Act was released and requires cities to decrease the concentration of air pollution.

This target-based performance evaluation has incentivized local officials to reduce air pollution. However, reducing air pollution is time consuming and costly. Researchers have argued that this results in "effortless perfection" through manipulation. The Ministry of Environmental Protection has also reported that there are mainly two types of manipulation: the first is called *soft manipulation*, which involves "modifying or manipulating the parameters and settings of the pollution reading equipment or software, or changing, deleting, adding, and falsifying the readings and data". The second type of manipulation is called *hard manipulation*, which involves "breaking the filter, adding, or removing parts of the reading equipment, such as adding diluting equipment to the filter intake vent, covering the reading machine, adding air filtering equipment next to the pollution monitoring station" (Xinhua Net, 2015). Regarding each type of manipulation, previous studies have proposed different statistical or econometric ways to identify these different types of manipulation, namely, manipulation at the cut-off, and systematic manipulation through shifting distribution. In the following sections, I discuss the sources of those different patterns of manipulation and anecdotal evidence that suggests them. In particular, the proportion of days with pollutants below the "Unhealthy" level is an important environmental target linked to the performance evaluation of local officials.

To sum up, decentralization opens the door to opportunism and hinders the independence of environmental departments. As a result, instances of local government interference in statistical departments have been widely reported and quantified (Ghanem & Zhang, 2014). Once local manipulation is detected, the environmental ministry will announce the manipulation and the director of the local environmental agency will be held responsible and face an administrative penalty. Based on empirical evidence and reports, two types of data manipulation happen in the decentralized system: *soft misreporting* and *hard misreporting*. The next section describes these two types of manipulation.

3 Pollution Misreporting

3.1 Soft misreporting

The first type of manipulation is soft misreporting. This type of pollution misreporting happens after the data is collection, in particular at the critical threshold specified by the promotion incentive. China has utilized a unique approach of regular performance evaluation and promotion

incentives to induce its local officials to comply with centrally mandated environmental targets. From 2012 to 2020, the Chinese central government used the number of days with unhealthy pollution levels as one of the performance metrics to evaluate the environmental achievements of local officials. The 12th and 13th Five-Year Plans (2010–2015 and 2016–2020), which are the most important national guidelines on overall development goals and road maps, set specific targets for different areas the annual proportion of days that are classified as healthy air quality by a national standard. Researchers argue that this naturally leads to an incentive to manipulate air pollution to be less than the threshold because such manipulation is hard to be noticed by the public. Previous studies have shown that this creates a strong incentive for local government to misreport at those thresholds which leads to discontinuity at a certain point over the distribution of the pollution (Ghanem & Zhang, 2014). In China, if days with good air quality make up at least 80% of days in a year, the city officials will be rewarded. In the study of Ghanem & Zhang (2014), they argue that the most likely form of manipulation happens on days where the pollution level is right above the threshold. The study finds significant clustering below the threshold, indicating that local data might be underreported to pass the environmental evaluation.

3.2 Hard misreporting

The second type of misreporting is hard misreporting. This type of manipulation interferes with the pollution monitoring station (and its surroundings) that collects air pollution data. This is largely motivated by the central government's requirement to reduce the overall air pollution level. In the 2014 Air Pollution Prevention Action Plan (In Chinese 大气污染防治行动计划), it states that in 2017, "all prefectural city PM_{10} levels should be 10% lower compared to those in 2012. "The Beijing, Tianjin, and Hebei regions, as well as the Yangtze River Delta, Zhu River Delta, and other regions, should reduce PM_{10} by 25%, 20%, and 15%, respectively". Similarly, in the 12th Five Year Plan 2010-2015, it specifies that by the end of 2017 there should be a 10% reduction in overall SO2 compared to 2010. The plan also specifies that in 2015, SO2, NO2, and PM 2.5 should be 10%, 7%, and 5% lower than in 2010, respectively.

Motivated by these air pollution reduction targets, the local government has been accused of tampering with monitoring stations in an attempt to lower the overall air pollution level in the data. This is usually done by breaking or interfering with the monitoring facility and equipment. For example, in December 2017, the Yulin Environmental Protection Department in Guangxi

Province installed automatic water spraying equipment around the monitoring station and planted trees and bushes around the station (Ministry of Ecology and Environment of China, 2018). In 2017, Ji'an city in Jiangxi Province was also found to have sprayed water on an air pollution monitoring station located in Hongsheng Factory (Ministry of Ecology and Environment of China, 2018). In January 2018, the national environmental protection bureau announced that multiple stations in Honghe, Yunnan Province, Xiangfan, Hubei Province, and other seven cities from 6 different provinces have "sprayed water on monitoring stations". In February, the bureau again announced two cities in Jiangxi and Henan province were charged with interference with the monitoring station readings through spraying water on the equipment (Xinhua Net, 2018)

A related type of hard misreporting, as suggested by the news report and anecdotal evidence, happens on the right tail of polluted days, i.e., days with high level of air pollutions. Several cities have reported having interference with the station during heavily polluted days. From the Ministry of Environmental Protection's Archives of Court Records, I discovered some publicly announced misreporting cases. On Aug 28th, 2018, a report named "Report on Linfen, Shanxi misreporting air pollution monitoring records" described the misreporting in detail, and I translated the report below.

"Verified by the local police department, in March 2017, in order to reduce the monitored air pollution level of Linfen City, the head of Linfen Environmental Protection Bureau, Zhang Wenqing, demanded the Director of Executive Office Zhang Ye, and Linfen environmental monitoring staff Yongpeng Zhang to interfering with six national controlled air quality monitoring facilities. Zhang Yongpeng pays 3,000 Yuan/month to 11 people involved in tampering with the station monitoring filter[8]. When Zhang Wenqing observed high air pollution, he told Zhang Yongpeng to "reduce the data". Zhang Yongpeng then told the 11 people in the WeChat group to tamper with the monitoring station. The criminals then spray water or sodium hydroxide on the intake sampler or cover the intake sampler, especially the analyzing equipment for $PM_{2.5}$, PM_{10}, and SO2. To avoid being taped by the surveillance camera, Zhang Yongpeng also pays 16,000 Yuan[9] to the monitoring staff to remove the footage of the surveillance camera during the interference." (Ministry of Ecology and Environment of China, 2018).

[8] 3000 Yuan is approximately $500 as of 2017.

[9] 16,000 yuan is approximately $2,500 as of 2017.

4 The environmental centralization reform

Falsification of pollution data can have a serious negative impact on public policymaking and is likely to result in a loss of social welfare. One example is the location choice of polluting firms. In China, when a heavily polluted city fails to improve environmental quality to a required level in a given period, it will not be able to obtain permits to build new polluting factories (Eaton & Kostka, 2014). If the local government underreports pollution levels to obtain a permit and attract more investment, then the residents will bear the welfare loss from excessive pollution emissions.

Underreported pollution may also lead to underinvestment in preventative health supplies. Studies have shown that short- and long-term exposure to $PM_{2.5}$ and PM_{10} has strong associations with adverse health effects (Dockery et al. (1993), Brunekreef and Holgate (2002), Gent et al. (2003), Bell et al. (2007)) such as ascending mortality rates, and morbidities such as a variety of cardiovascular diseases (Ramanathan et al. (2001), Lin et al. (2002), Gauderman et al. (2004), Dominici et al. (2006)). In order for people to be fully cautious about their exposure levels to PM, truthful determination of pollutant levels is essential.

Finally, data manipulation incurs additional verification and inspection cost. Since the misreporting happens because the decentralized environmental management system in China results in local government having too much informational advantage, other channels are built to bridge the informational gap between the local and upper-level government. For example, if the central government is concerned about the quality of reports submitted by the local government, the central government dispatches inspection teams to all provinces across China to scrutinize the actual quality of the environment (Zheng & Na, 2020), incurring additional inspection costs and resources spent. (Zheng & Na, 2020)

The central government has acknowledged this misreporting issue and has implemented a series of measures to reinforce the oversight of the central government over local authorities (Brombal, 2017). More technical approaches have been adopted to increase the difficulty of manipulating data. There have been more frequent announcements of violations, and the statistics law has been revised to address the punishment of data fraud. Specifically, the previous Statistics Law in China has been in effect since 1983, but it was too vague to enforce. Although it stated the penalties for illegal acts, the law did not specify the types of illegal acts and the extent to which penalties should be imposed. In 2008, China passed a new Statistics Law. This law lists four types

of statistics cheating: revising statistics without permission or making up statistics; forcing or ordering statistics departments or individuals to revise or make up statistics or refuse to report statistics; retaliation against individuals who refuse to issue false statistics; and retaliation against individuals who report statistics violations. The degree of punishment depends on the consequences of the violations, and the punishments include a warning, recording a demerit, or even removing officials from their positions.

To further reduce pollution misreporting, in November 2015, a draft of environmental vertical reform was proposed. In July 2016, the official document was released to the public. The official name of the reform is "Guideline of Pilot Program of the Vertical Management Reform for the Monitoring, Supervision and Law Enforcement of Environmental Protection Agencies below the Provincial Level.". According to the national guidelines, the document was formally effective in October 2016 and is known as the Environmental Vertical Reform. Participation is voluntary. Hebei, Chongqing, Jiangsu, Shandong, Hubei, Qinghai, Shanghai, Fujian, Shaanxi, Jiangxi, Tianjin, and Guangdong provinces participated. Once a province has enrolled, all cities in that province will automatically participate in the centralization reform.

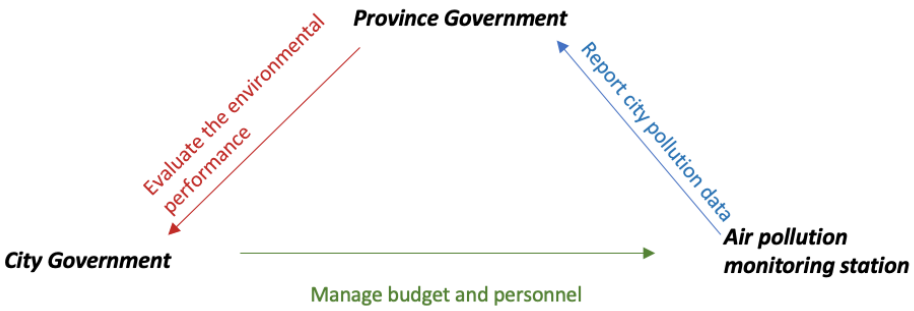


Figure 3-1a Structure of environmental agency before the reform

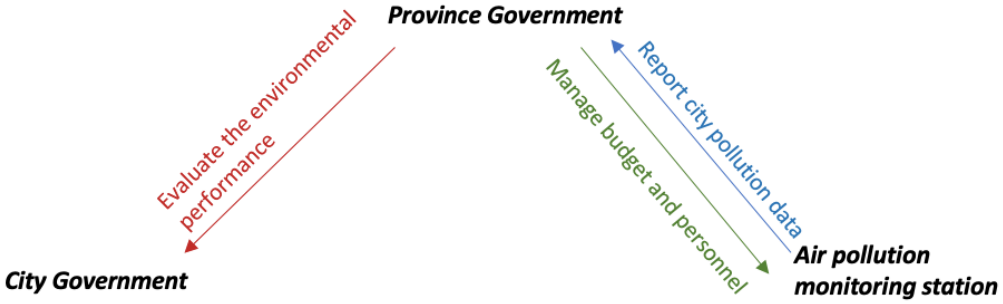


Figure 3-1b Structure of environmental agency after the reform[10]

Figure 3-1 Structure of environmental agency before and after the reform

This reform has two notable features. First, it does not alter the political incentives for reducing air pollution, so the incentive to reduce pollution remains. The second is the adoption of a vertical structure (Figure 3-1, bottom figure) whereby an agency works via an internal hierarchical structure, with lower functional units reporting directly to upper ones instead of to territorial governments (Ma Y. , 2017). After the reform, the City Environmental Protection Bureau remains as a functioning department of the city government. This reform is believed to be a fundamental change to the old environmental governance structure by decoupling environment-related authorities from local interests. As a key step to ensuring the reliability of environmental statistics, this reform is designed to add institutional barriers to pollution misreporting.

However, the reform only breaks the formal bond between local interests and environmental agencies. Whether it can also decouple the informal bond remains uncertain.

[10] The green arrow showing the change in the structure is based on what is stipulated in the official document of the reform.

Moreover, the reform introduces heterogeneity into the system. Untreated governments would presumably face the closer examination mentioned earlier, since they might be considered more vulnerable to data manipulation. In fact, after the reform, data manipulation was still detected even in some cities that participated in the reform, suggesting that the reform did not eradicate the manipulability of pollution records (South China Morning Post, 2018). Given this, the null hypothesis is: There is no statistically significant change in the difference in air quality data manipulation between cities that are involved in the vertical reform and those that are not.

5 Conceptual framework

In this section, I use a simple model to describe how the centralization reform, which is assumed to have increased the misreporting cost parameter, reduces equilibrium misreporting. I start by formalizing the comparative statics in a generic model. Suppose local government maximize benefit from reporting $B(R)$, minus cost from misreporting. I denote misreporting as Δ, or the difference between the true air quality Y and the reported air quality R. The higher the reported air quality, the more benefit as local government is rewarded $B'(R) \geq 0$; the higher the gap between true quality and reported quality (misreporting), the higher the misreporting cost $C'(\Delta) > 0$. For simplicity I also assume that the benefit is linear such that $B'(\Delta) = 0$, and I assume that it is separable and additive in true pollution and reported pollution, and that the cost is convex in the size of misreporting, $C''(\Delta) > 0$.

The local government maximize the objective function as follow

$$\max B(R) - aC(Y - R)$$
$$= \max B(Y + \Delta) - aC(\Delta)$$

With the additively separable assumption:

$$= \max B(Y) + B(\Delta) - aC(\Delta)$$

The first order condition with respect to Δ is

$$B'(\Delta) = aC'(\Delta)$$

With the reform, the marginal benefit remains unchanged while the marginal cost of misreporting a has increased. By implicit function theorem $\Delta'(a) = -\frac{C''_{\Delta a}(\Delta(a),a)}{C''_{\Delta\Delta}(\Delta(a),a)} < 0$. In the equilibrium, an increase in the marginal cost of misreporting reduces misreporting. This comparative statics is visualized in Figure 2. The solid blue line represents the constant marginal

benefit of misreporting, and the solid black line represents the marginal cost of misreporting before the reform. Everything else remains constant in the sense that performance evaluation standards continue to reward pollution abatement, and pollution abatement policymaking is still delegated to local governments. Following the reform, there is more control by the central authority over information on performance, making misreporting more difficult (and costly) for the agent. The reform raises the marginal cost of misreporting for the agent, thereby reducing privately optimal misreporting (Delta prime is smaller than Delta), as shown in Figure 3-2.

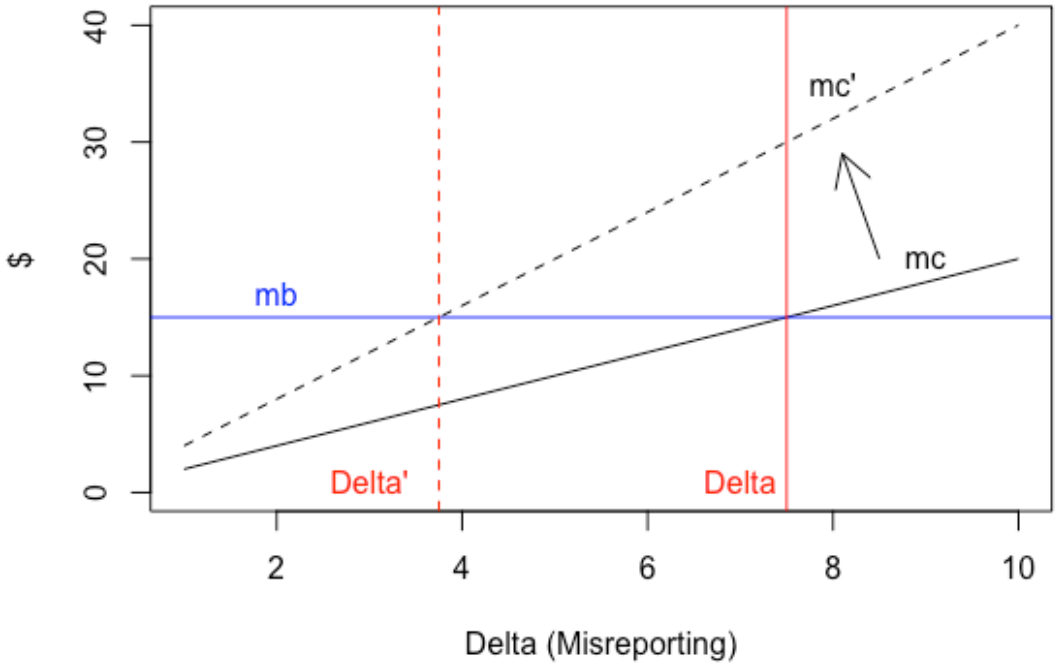


Figure 3-2 Comparative static before and after the reform

The comparative statics predicts that increase in marginal misreporting cost will lead to reduced misreporting. It does not inform which type of misreporting it reduces given the context in China. Therefore, I combined the insights from the comparative statics with the institutional detail in China. Figure 3-3 shows conceptually how this centralization reform might affect local misreporting, both soft and hard misreporting, with examples, respectively.

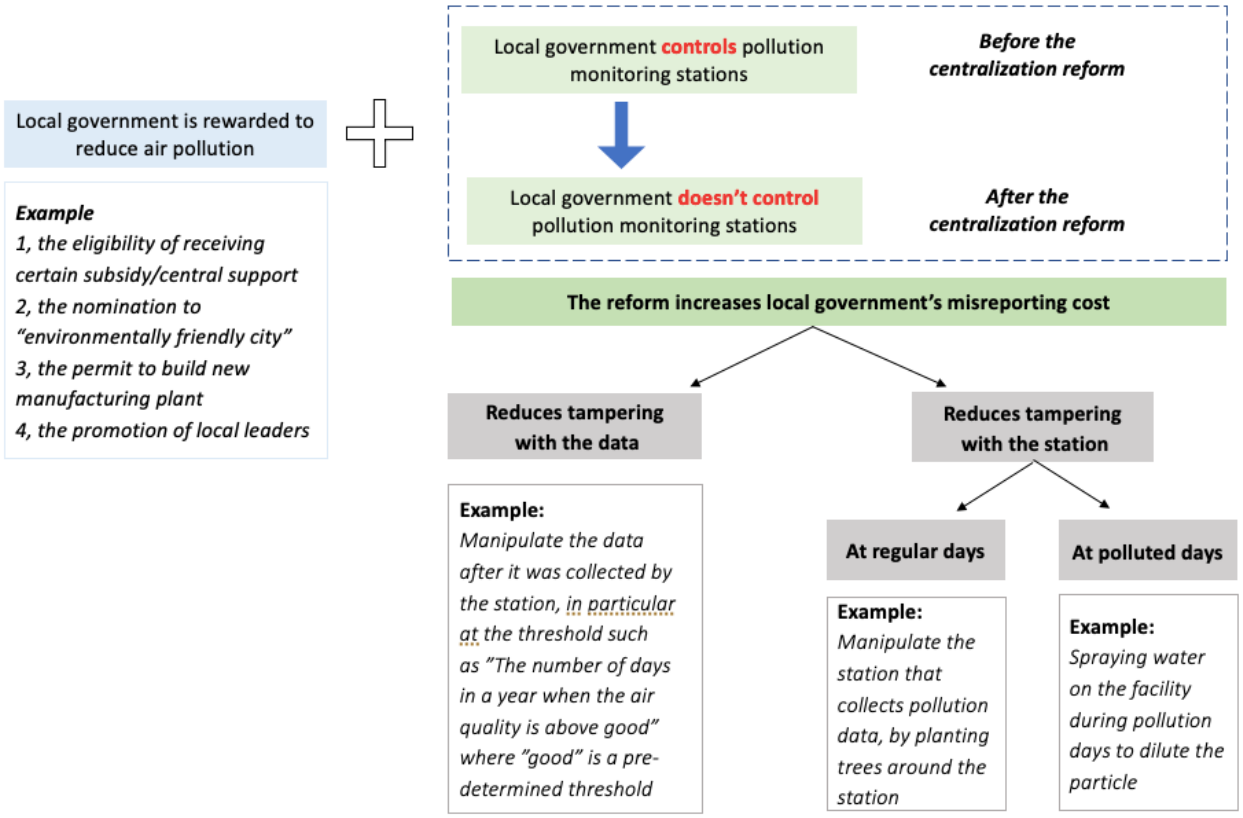


Figure 3-3 Conceptual framework of the reform

6 Data

6.1 Treatment variable: Reform participation

The treatment variable is reform participation. The reform was formally effective starting in October 2016 and is known as the Environmental Vertical Reform. Participation is voluntary. In total, among all the 32 provinces in China, 12 provinces participated in the reform. The participating provinces are Hebei, Chongqing, Jiangsu, Shandong, Hubei, Qinghai, Shanghai, Fujian, Shaanxi, Jiangxi, Tianjin, and Guangdong provinces. Once a province has enrolled, all cities in that province will automatically participate in the centralization reform. The rest of the provinces in China are considered as control observations. The timing of the reform and the list of participating provinces were obtained from the Ministry of Ecology and Environment and Xinhua.net. Since each province is allowed to set its own effective date, I further obtained this information from the government website of each province and news reports.

6.2 China official air pollution data

I study the misreporting of air pollution using data from the Chinese government. I have obtained two disaggregated air pollution datasets. The first is monitor-hourly data for PM_{10} and $PM_{2.5}$ obtained from the Ministry of Environmental Protection. PM_{10} and $PM_{2.5}$ are considered major air pollutants in China. The data spans from June 2014 to September 2018. The number of monitoring stations has increased gradually from 998 in 2014 to 1,600 in 2018. The location of the monitoring station is specified by the National Environmental Monitoring Department. I geocode the exact location of each monitoring station using their coordinates. For all the analysis, I only use information from those 998 monitoring stations that are available for the entire study period. Figure 3-4 plots the location of the monitoring stations.

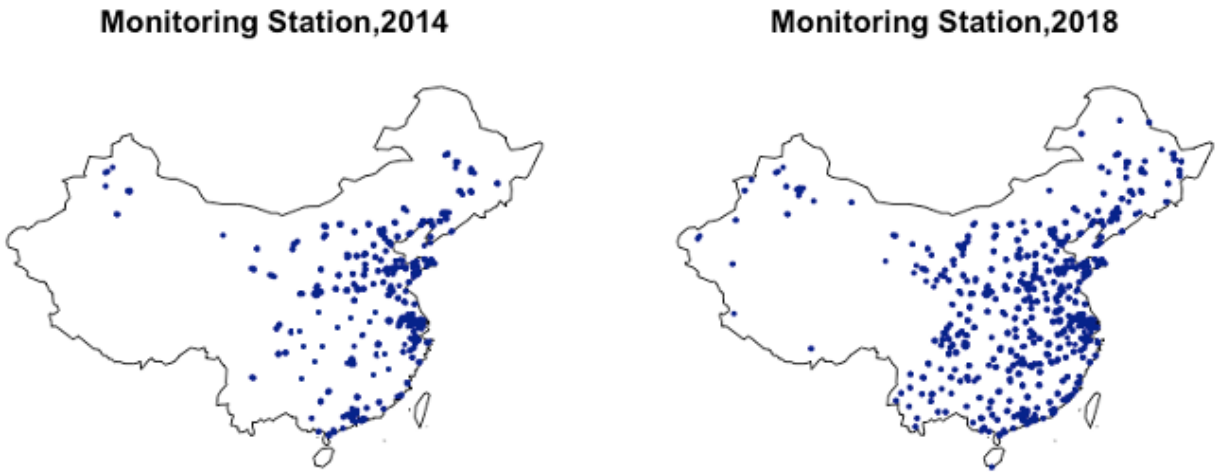


Figure 3-4 Location of the monitoring station

6.3 NASA MODIS Terra satellite AOD

The second set of data is NASA's monthly AOD (Aerosol Optical Depth) raster data. This data will be used as the benchmark of the true air pollution level, after controlling for weather conditions and other variables. The AOD data is retrieved from remote sensors known as the Moderate Resolution Imaging Spectroradiometer (MODIS) Terra and Aqua satellites (NASA 2010). The AOD captures the amount of radiation absorbed, reflected, and scattered due to the presence of solid and liquid particulates suspended in the atmosphere (Chen et al 2013).

Researchers have shown that the AOD, corrected for meteorological conditions, can predict group level PM (Gupta et al. 2006; Kumar 2010; Kumar et al. 2011) and scholars have shown that AOD captured 70% of the variations in the PM_{10} monitored on the surface after controlling for meteorological conditions.

The monthly AOD data was downloaded at a 1 km spatial resolution. AOD is potentially available everywhere at the satellite crossing time (10:30 am and 1:30 pm of Beijing time) but with many missing data points. Studies have shown that it is sensitive to point-specific and time-specific weather. It is only available on days when there is less than 10% cloud cover. The missing data could also be due to satellite sensors experiencing difficulties in retrieving AOD data in arid and semi-arid regions that have a bright background. As a result, I use the AOD data every month to ensure that there are enough observations for each grid. For this study, I removed all the pixels that were missing data. To get a station-level average AOD, I average AOD across all pixels within a 10 km radius of each pollution station mentioned above, and this number ranges from zero (clearly) to one (heavy coverage). Figure 3-5 shows an example of AOD in one month.

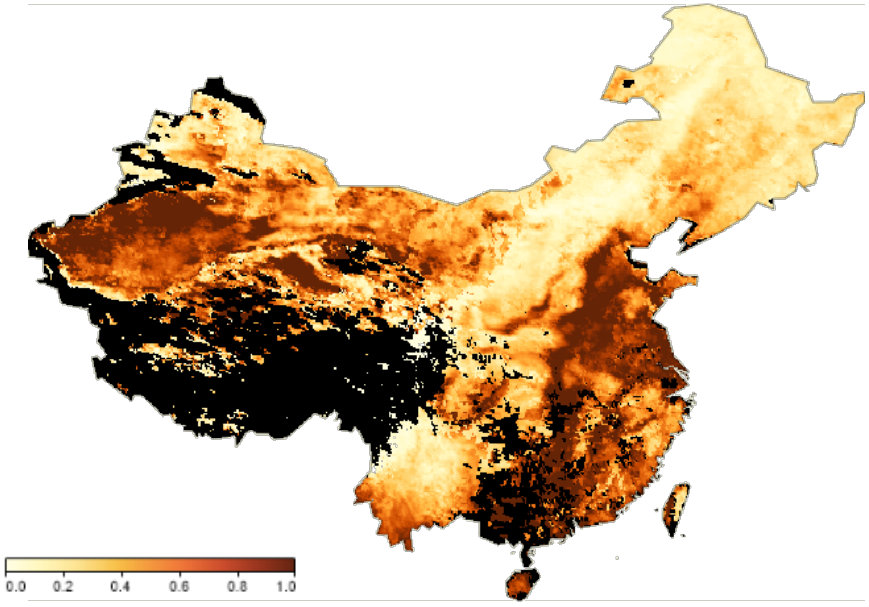


Figure 3-5 Example of monthly Aerosol Optical Depth

Note: NASA Terra Modis on April 2018. Areas with black colors represent missing AOD values due to heavy cloud cover or bright background such as dissert areas

Figure 3-6 shows the overall trend in AOD from 2000 to 2018. By visual examination, the (unconditional) AOD level seem to have been decreasing since 2012, and the overall negative trend is statistically significant at 10% level. Since AOD is a proxy for PM_{10}, this trend indicates that the true pollution in China from is likely to have decreased, especially in recent years.

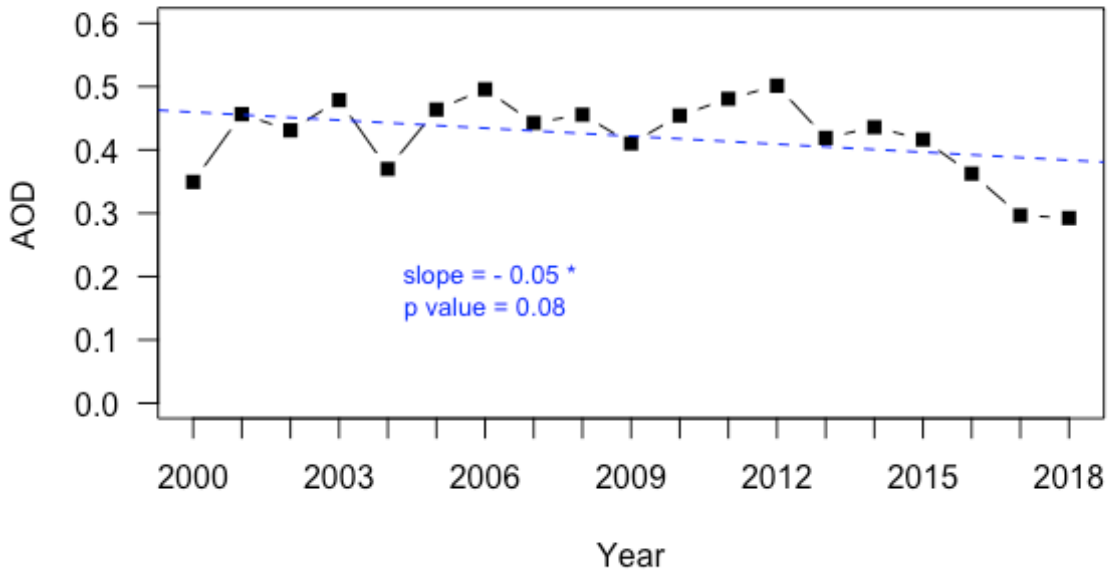


Figure 3-6 AOD trend 2000-2018

6.4 Weather, demographic, and economic data

Aside from air pollution data, control variables will be obtained from the China City Economic Database, and weather data from the National Metrology Center. These control variables are used to reduce the variance in air pollution measurements. The weather data includes daily air temperature, dew point temperature, pressure, 6-hour precipitation, wind speed, and sky cover. The dew point temperature is the temperature to which the air must be cooled to become saturated with water vapor. It is widely used to measure humidity as a higher dew point means there is more moisture in the air, which tends to dampen air quality. The sky cover measures the Horizontal Infrared Radiation Intensity, which measures cloudiness. Each weather station has a corresponding longitude and latitude. To find the average weather condition for each pollution monitoring station, I draw a circle with a 10 km radius around each station and calculate the average weather condition. Figure 3-7 shows the location of the weather stations.

Figure 3-7 Location of weather stations

Table 3-1 shows the summary statistics of each variable. For air pollution information, I have more than 47,000 station-month observations. Among them, more than 25,000 stations have weather information, and more than 56,000 can be linked to the monthly average of the AOD.

Table 3-1 Summary statistics

	Unit	Obs.	min	max	mean	std.dev
AQI	NA	47,904	14.98	381.91	75.15	33.08
PM_{10}	u/m3	47,904	4.80	1539.77	84.63	46.97
$PM_{2.5}$	u/m3	47,904	3.65	653.23	47.67	27.41
Air Temperature	Celsius	25,734	-22.18	32.07	15.95	10.70
Dew Point Temperature	Celsius	25,734	-27.20	27.62	9.09	12.22
Pressure	Hectopascals	24,525	998.58	1043.08	1015.13	85.09
Wind Speed	m/second	25,734	0.64	6.59	2.66	0.77
Sky Cover	W/m2	25,726	0.15	9.00	5.58	1.49
Precipitation (6 Hours)	mm	24,957	0.00	237.33	33.15	27.22
AOD	Unit free	56,419	0.00	1.00	0.45	0.20

7. Identification and Estimation Strategies

I start by describing how to quantify data manipulation. There are two common empirical approaches to detecting data that has been misreported.

The first is to test statistical patterns that are unlikely to exist without manipulation. For example, Benford (1938) tests falsified financial data relying on the statistical derivation of digits

frequency. He shows that certain digits should appear less/more frequently than others. This method has also been widely adopted to test manipulation, especially in income and taxation (Stoerk, 2016), yet it applied mainly to cases where the number is generated through certain distribution and the numbers must span multiple orders of magnitude (e.g., data that ranges from 100 to 10,000,000). It is less suitable to apply this method to circumstances like air pollution or economic data (Miranda-Zanetti, 2019), and more recently, it has been applicated in exploring potential data manipulation in election and COVID cases but in both cases it is not yet clear if Benford law is appropriate (Reuters, 2020). Later, McCrary (2008) relies on a nonparametric test of density at different values of data points, and the assumption is that the data should have no difference in density around pre-defined cut-offs without manipulation.

The second approach is to test data patterns that are consistent with manipulation, patterns that are very unlikely to exist when there is no manipulation. Ghanem & Zhang (2014) investigate discontinuities at artificial cut-offs across different levels of air pollution. This is essentially a Regression Discontinuity Design (RDD) method in order to measure the magnitude of manipulation. Their approach was later extended to measure not only existence, but also the magnitude of manipulation (Ghanem, Shen, & Zhang, 2020). Fisman & Wang (2017) study a program designed to reduce accidental deaths and find sharp discontinuity in reported deaths at the death ceiling, suggestive of manipulation; Kalgin (2016) compares the reported indices that should follow a normal distribution and quantifies the misreporting using the deviation from a normal distribution. Acemoglu et al (2020) investigate the Columbian colonel's misreporting behavior when confronted with a high power. This data is unlikely to suffer from the systematic biases of estimates from official sources and victim associations. Martinez (2018) uses nighttime light outer space data as an unbiased approximation of true economic activity and compares official GDP with this dataset to determine the magnitude of manipulation.

In this paper, I will use both approaches because they are complements to the two types of manipulation mentioned earlier. The McCrary test is able to identify discontinuity that results from soft misreporting. Using NASA satellite data allows me to identify the hard misreporting that deviates continuously from the true pollution distribution. In addition to hard misreporting during regular days, I also extend it to studying the potential misreporting at the tail of the air pollution data, i.e., during heavy polluted days.

7.1 Estimating soft misreporting

In the first approach, I adopt the discontinuity test to identify if the collected data had been manipulated. This approach utilizes the fact that in China, environmental performance evaluation by the local government is based on pre-defined thresholds. The estimation involves two steps.

In the first step, I run more than 1,600 RDD tests for each province each month. This RDD method to test data manipulation was initially designed by McCrary (2008) and later applied by Ghanem & Zhang (2014) in the context of China's air pollution. The key assumption is that (1) the density function for pollutant concentrations is continuous and (2) that regulators have imprecise control of the air pollution around the cut-off. I use hourly PM_{10} data for each province each month. More specifically, there are 53 months in total, with 8 months for 2014, 9 months for 2018, and 12 months for the rest of the years. This gives around 720 observations to conduct the McCrary test for each province each month. As previously stated, the manipulation incentive occurs at 150 for PM_{10}, with anything above 150 considered unhealthy. Using data-driven bandwidth, in total, there are 1,643 province-month McCrary test results. This generates a variable *misreport* that equals 1 for months during which the province has significant McCrary test results based on one-side test criteria.

In the second step, I estimate the effect of reform participation on the McCrary test results from the first step in a binary logit regression. The outcome variable is *misreport*. This variable is from the first stage and has a value of 1 if the month is identified as having questionable pollution data. The treatment variable is whether the city participates in the reform program. The treatment indicator is then interacted with the post-treatment period dummy as in a standard DID estimation. I included the fixed effect of the month to control for seasonality and the fixed effect of the province to control for province-level characteristics that might correlate with the outcome. Thus, the model extends the DID into a more general form as follows.

$$\frac{Log\left(Prob(Misreport_{it}=1)\right)}{1-Prob(Misreport_{it}=1)} = \beta_0 + \beta_1 D_i T_t + \gamma X_{it} + province_i + month_t + \varepsilon_{it} \quad (1)$$

Where $Misreport_{it} = 1$ is whether province i month t shows a significant discontinuity at the threshold, D_i is a dummy variable that equals 1 if province i is treated, T_t is a dummy variable that equals 1 for post-treatment periods. $\boldsymbol{X}_{it}$ is a vector of control variables. The

coefficient of interest is β_1. If estimated β_1 is negative, it indicates that joining the reform reduces the statistically detected manipulation. The identification assumption is the exogeneity of treatment assignment. Further discussion of this assumption is provided later in the robustness check section.

7.2 Estimating hard misreporting

For the second type of misreporting, or hard misreporting, I estimate the abnormal patterns that happen at the station level. I use the NASA AOD data as an unbiased benchmark of true pollution levels. The ability to map growth in NASA data to growth in China's official air pollution data provides an estimate of the magnitude of manipulation. The specification largely follows the one used in Martinez (2018). The identification assumption is that the reform participation is as good as random once I control all the right-hand variables. I use the growth rate in the baseline specification. Transforming the level variable to the growth rate in a regression model is a common way to handle situations where non-linear relationships and because the absolute level of AOD and PM pollution are of very different scales: AOD ranges from 0 to 1 while the PM ranges from 10 to 500. On the other hand, the growth rate relationship is easier to compare. The estimation follows the specifications below:

$$growth\ rate\ PM_{10_{it}} = \beta_0 + \beta_1 growth\ rate\ AOD_{it} + \boldsymbol{\beta_2 D_i T_t growth\ rate\ AOD_{it}} + growth\ rate\ AOD * X_{it}\gamma + province_i + month_t + \varepsilon_{it} \quad (2)$$

Where $growth\ rate\ PM_{10_{it}}$ is the monthly growth rate of China's reported PM_{10} in station i month t, $growth\ rate\ AOD_{it}$ is the monthly growth rate of AOD (true pollution benchmark) in station i month t, D_i is a dummy that equals 1 for the treated station, T_t is the dummy that equals 1 for post-reform periods. β_1 is interpreted as AOD elasticity of PM_{10}, and β_2 is interpreted as how joining the reform affecting the mapping from AOD data to official PM_{10} (or the reform gradient in the AOD elasticity of PM_{10}). As mentioned in Martinez (2018), this implicitly assumes proportional misreporting.

Without manipulation, the mapping from NASA AOD records to official records should be the same across the treated and control stations. If, instead, $\hat{\beta}_2 > 0$, then there is evidence that treated stations have a higher elasticity of NASA AOD data, suggesting less manipulation. This equation is a generalized form of DID, which can control for omitted variables that are time-constant, such as city-specific local-central relationship, or public preferences over environmental quality. In a two-way fixed effect DID, station-specific effects account for time-invariant confounders that are specific to each station. Month fixed effects can control for seasonality. In addition, since the AOD data is obtained as raster while the data from the pollution station is spatial points, for each pollution monitoring station, I draw a 10 km (6.2 miles) radius around the station and average the Aerosol Optical Depth for each station each month.

The covariate matrix X represents a set of weather variables and social-economic variables. Meteorologic literature has widely used AOD in predicting ground-level PM_{10} and $PM_{2.5}$. Prediction of ground-level PM by AOD data is affected by several factors (Kaufman and Fraser, 1983; Remer et al., 2005; Levy et al., 2007; Liu et al., 2008). For example, temperature (measured by air temperature) and humidity (measured by dew point temperature) can affect particle formation rate through photochemical oxidation and condensation process imposing different particle composition. Wind speed can affect the AOD and PM_{10} relationship as greater wind speed would dilute the concentration of pollutants or transferring particulate matter from different sources. Moreover, change in the vertical profile of PM and AOD due to the cloud contamination in the upper air could affect the PM-AOD relationships. Therefore, I also include the sky coverage measure as a covariate.

Notice that the equation above is a standard linear panel data model devoid of spatial effects. This model can be used as a reference for the estimation results of spatial panel data models as well as to check the robustness of these estimation results (Yang et al., 2017). One limitation of Equation (1) is that it does not consider the spatial autocorrelation of air pollution and the spatial correlation between air pollution and other unobservable factors in neighboring cities. Air pollution transport is one of the most dynamic atmospheric processes, as air pollution emissions can easily impact the air of neighboring regions through the wind. Under certain weather and geographical conditions, local air pollution can spread to a wider area. Failing to incorporate this will result in biased estimation of the coefficients.

I conduct Moran's I test on the following air pollutant using distance matric. Figure 3-8 shows significant spatial autocorrelation as most of the points are located on the line with a positive slope, suggesting both AOD and PM_{10} are positively correlated across space. The size of the correlation is consistent with previous literature.

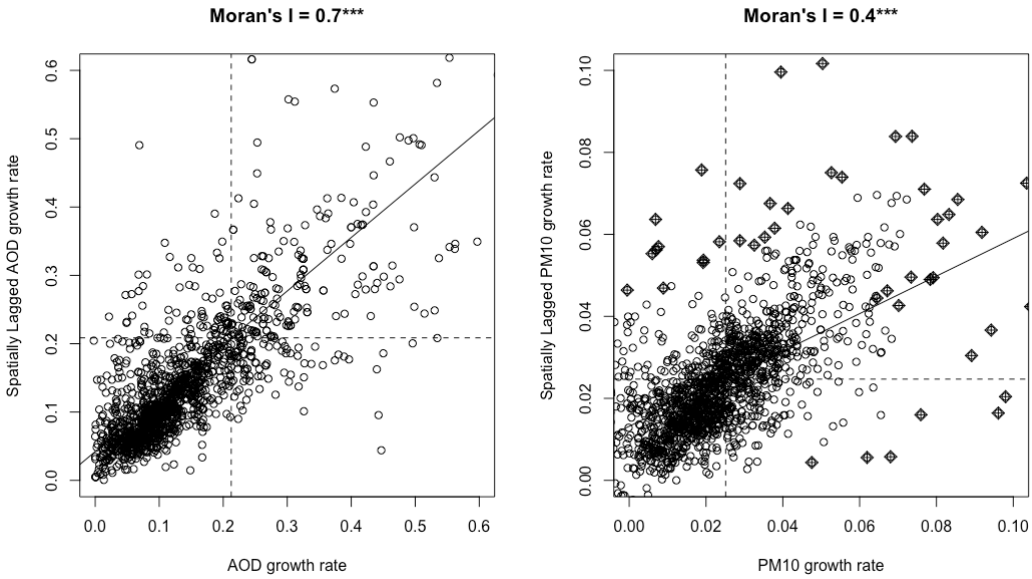


Figure 3-8 Moran's I for aggregate AOD and PM_{10} 2014 – 2018

Motivated by the Moran's I result, I now turn to the spatial panel model. I experimented with two ways to incorporate these spatial effects into the regression model: the spatial-lag model to capture the spatial dependence of air pollution, and the spatial lag of the error term. The spatial-lag model implicitly assumes that the spatially weighted average of air quality in a neighborhood affects the air quality in addition to the other explanatory variables. I also consider the above model with a spatial autoregressive component. The spatial lag model comes from the fact that air pollution, especially fine particles like PM_{10}, are highly spatially correlated. Pollution in one city can affect the air quality in adjacent cities. Therefore, a spatial lag model is defined as follows:

$$growth\ rate\ PM10_{it} = \alpha + \beta_0 growth\ rate\ AOD_{it} + \rho W growth\ rate\ PM10_{it} + \beta_1 \boldsymbol{D_i T_t growth\ rate\ AOD_{it}} + X_{it}\gamma + province_i + month_t + u_{it}, u_{it} = \lambda W u_{it} + \varepsilon_{it}, \varepsilon_{it} \sim N(0, \sigma^2 I_n) \quad (3)$$

where ρ is the spatial autocorrelation parameter, λ is the spatial autocorrelation of the error term, W is a spatial weight matrix (which will be justified in the next section). This also implies that air

pollution is considered to be a global effect rather than a local one, as the spatial multiplier $(\mathrm{I}-\rho\mathrm{W})^{-1}$ shows up in the marginal effect expression (See Anselin, 2003):

$$growth\ rate\ PM10_{it} = (\mathrm{I}-\rho\mathrm{W})^{-1}[\alpha + \beta_0 growth\ rate\ AOD_{it} + \rho W growth\ rate\ PM10_{it} + \beta_1 \boldsymbol{D_i T_t growth\ rate\ AOD_{it}} + X_{it}\gamma + province_i + month_t + (\mathrm{I}-\lambda W)^{-1}u_{it}] \quad (4)$$

This model, with both a spatially lagged dependent variable and spatially lagged error term, is labeled by Anselin (1988) as a *SARAR* model (Spatial Auto-Regressive model with Auto-Regressive disturbances, or SARAR for short). The introduction of spatial lag also results in an endogenous variable $\rho W y$. As OLS is biased with the spatial lag model, a maximum likelihood estimator is used. By contrast with OLS, MLE have attractive asymptotic properties, which apply in the presence of spatially lagged terms.

Because the unit of analysis in the dataset is point data, I use k nearest neighbor spatial neighbor list. I choose 5 nearest neighbors. Formally, let the distances from each spatial unit i to all units $j \neq i$ be ranked in descending order as follows: $d_{ij(1)} \leq d_{ij(2)} \leq d_{ij}$ Then for each $\mathrm{k} = 1, \ldots, \mathrm{n}-1$, the set $N_k(i) = \{\mathrm{j}(1), \mathrm{j}(2), \ldots, \mathrm{j}(\mathrm{k})\}$ contains the k closest units to i (where for simplicity I ignore ties). For each given k, the k-nearest neighbor weight matrix, W, then has spatial weights of the form: The corresponding spatial weights have the following form:

$$w_{ij} = \begin{cases} 1, & \mathrm{j} \in N_k(i) \\ 0, & \text{otherwise} \end{cases}$$

In addition to the regression on the mean, I also estimate the manipulation that happens at the tail using spatial quantile regression. In OLS estimation, the relationships between the outcomes of interest and the explanatory variables remain the same across different values of the variables. In our context, however, I am also interested in the effect of the reform across the distribution of pollution variables rather than only at its mean. Using quantile regression, I estimated the relationship between AOD and PM10 using quantile regression and evaluated the upper quantiles using the following specification in equation (5). I am most interested in the right tail, namely the 75, 85, 95% quantile where $\tau = 0.75, 0.85\ and\ 0.95$.

$$PM10_{it} = \beta(\tau)_0 + \rho W PM10_{it} + \beta(\tau)_1 D_i T_t AOD_{it} + X_{it}\gamma + province_i + month_t + \varepsilon_{it} \quad (5)$$

8 Results

8.1 Does the reform reduce soft misreporting?

To test if the reform reduces soft misreporting, I conduct a two-step estimation described in the previous sections. I first conduct McCrary Test for testing manipulation at the cut-off point (i.e., where the incentive is). The level of analysis is province-month. This gives us more than 1,600 RDD results indicating whether a province in a certain month is associated with statistically questionable air pollution. Next, I use the results from step 1 as the dependent variable, and test if such measurement of manipulation changes after the reform implementation using a standard DID model, controlling for relevant covariates, province fixed effect, and month fixed effect.

First stage: Estimates of discontinuity at the threshold

The first step is to conduct a McCrary test for each province each month. The null hypothesis for the McCrary test is that there is no density difference to the left and the right of the threshold. The cut-off point is where the manipulation incentive is. For PM_{10} this cut-off is 150, as the environmental performance evaluation stipulates that "the number of days with PM_{10} greater than 150 should be less than 200 days". Based on one-sided test criteria, I find that among all the 32 provinces, 10 have at least one month in the 5-year span that exhibits questionable pollution records, suggesting manipulation at the cut-off. Notice that having dubious reporting does not necessarily indicate manipulation, but rather indicates a statistical pattern that is less likely to occur with truthfully reported data. Figure 3-9 compares two provinces in the same month. One presents normal air pollution (left figure) and the other one shows a significant discontinuity at the threshold (right figure), implying possible dubious air pollution reporting.

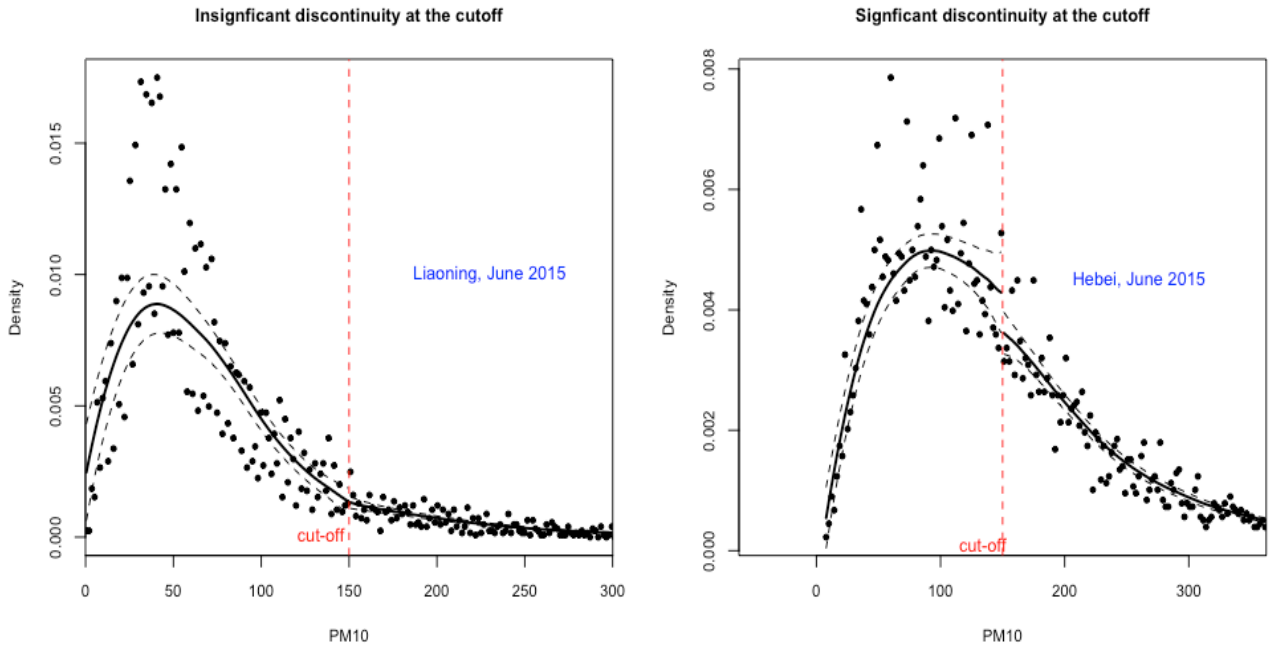


Figure 3-9 Selected provinces with insignificant and a significant discontinuity

After obtaining the estimated misreporting for treated and control provinces, I test for pre-existing trends. The pre-trends test is a common way of assessing the plausibility of the parallel trend assumption in the difference-in-differences designs. The DID assumes that the trends in misreporting in the absence of centralization reform in treated regions would not differ from the trends in non-treated regions. While I can not observe the misreporting in the absence of the reform for treatment provinces, I can compare treated and control regions during the periods prior to the reform implementation. I use only the pre-reform data and I regress misreporting on control variables. Figure 3-10 plots the test results to visually examine the pre-trends. The plot shows that the predicted misreporting follows a similar trend up until 8 months prior to the reform. Before the reform, the difference between the treated and control provinces had started to decrease. This might raise questions about the DID design's validity: it's possible that some factor was driving the treatment effect prior to the reform's implementation.

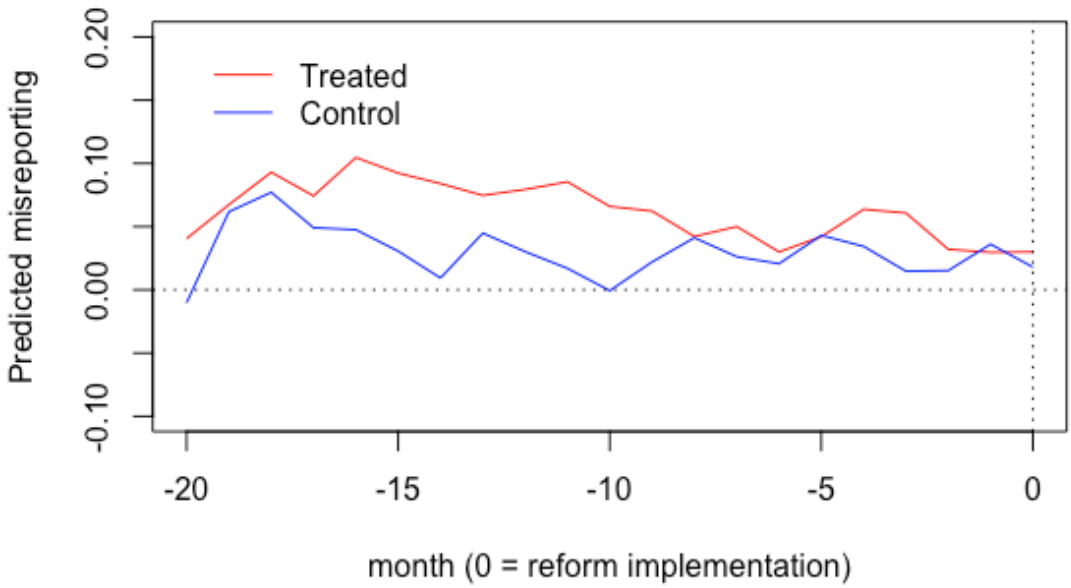


Figure 3-10 Pre-trend test for predicted misreporting (soft misreporting)

Second stage: effect of the reform on soft misreporting

In the second step, I use the results from the first stage RDD as the dependent variable in a standard DID model, and to test if the discontinuity at cut-off is different for treated provinces after the reform was implemented. Since the estimated results are used as the dependent variable, following Hornstein & Greene (2012), all observations are weighted by the inverse of the variance of the RDD results obtained from the first stage estimation. The second stage results are presented in the figure below. On average, before the reform, among all the provinces, 2.4 provinces seemed to have dubious PM_{10} reporting in a given month. After the reform, among all the provinces, 2.6 provinces seem to have dubious PM_{10} reporting in a given month. Overall, I do not find a significant change in this type of manipulation after the reform as the confidence interval for the treatment effect does not differ significantly from zero. Figure 3-11 presents the treatment effects by month. The reform month is shown by the vertical dashed line at 0.

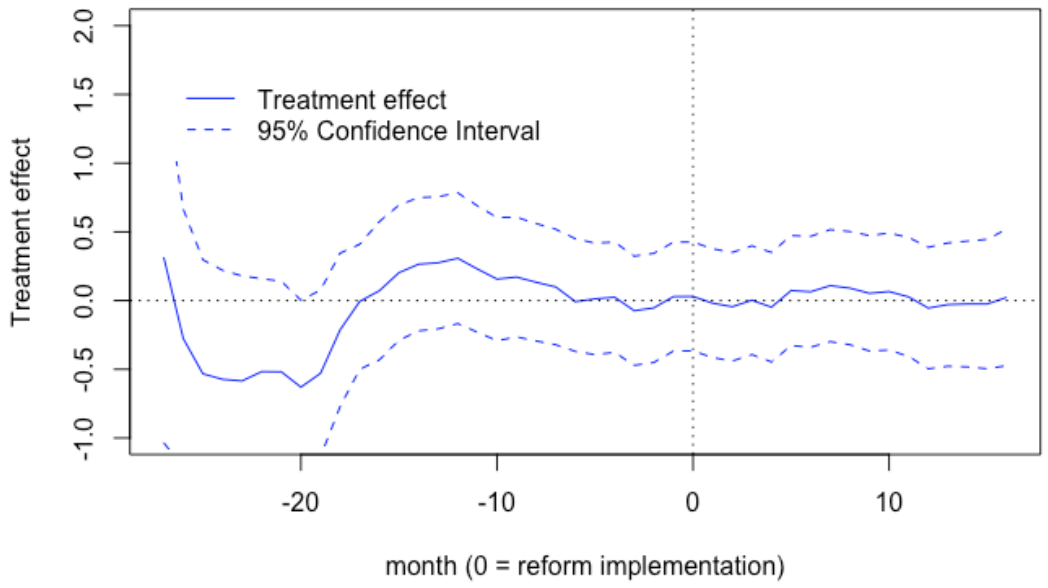


Figure 3-11 Treatment effect by month (soft misreporting)

One possible explanation is that the installation of automated pollution equipment reduces soft misreporting significantly. Although our results suggest that there is still some weak evidence of misreporting, the magnitude is very limited, and our model is not able to identify any significant reform impact on it. Previous study has explored the effect of the installation of automated pollution equipment. As suggested by Greenestone et al (2020), the installation of automated pollution monitoring has improved the technical difficulty of modifying or falsifying air pollution data. During the study period, automatic monitoring technology changed as well, resulting in an attenuated effect of the reform that we study. The automation of the national air quality monitoring network involves the establishment of a new real-time reporting system. Importantly, both the monitoring equipment and the method of measuring PM_{10} remain unchanged, ensuring that any differences in PM_{10} are not the result of changes in equipment or method. Instead, the existing equipment was integrated into the new monitoring system. The primary feature of the new approach to monitoring is real-time reporting, which enables online validation and higher-standard requirements for measurement.

After the installation, all the readings will be directly uploaded to city, province, and national environmental monitoring databases, also known as “three uploads” (People Daily, 2018). This makes soft misreporting much harder than before. The upgraded automatic monitoring system was installed in different cities in three waves. In the first wave, 74 cities (with 496 stations) were

upgraded by January 2013. In the second wave, another 116 cities (with 449 stations) were upgraded by January 2014. In the final wave, 177 cities (with 552 stations) were upgraded by November 2014. Because these installations take place at least two years before the centralization reform, there is less concern that the upgrading will interfere with the reform's effect.

Given the study period of this paper, Figure 3-12 compares the trend of reported pollution and the proxy for true pollution before and after the third wave of the installation which was December 2014. The y axis on the left (the black axis) shows reported PM_{10}. The y axis on the right (the blue axis) shows the AOD data. I find that right after the third wave of the installation, there is a sudden increase in the reported pollution. This increase is significant at 1% level. On the other hand, there is no change in the official PM_{10} for cities that have already installed automated monitors and no change in the NASA data.

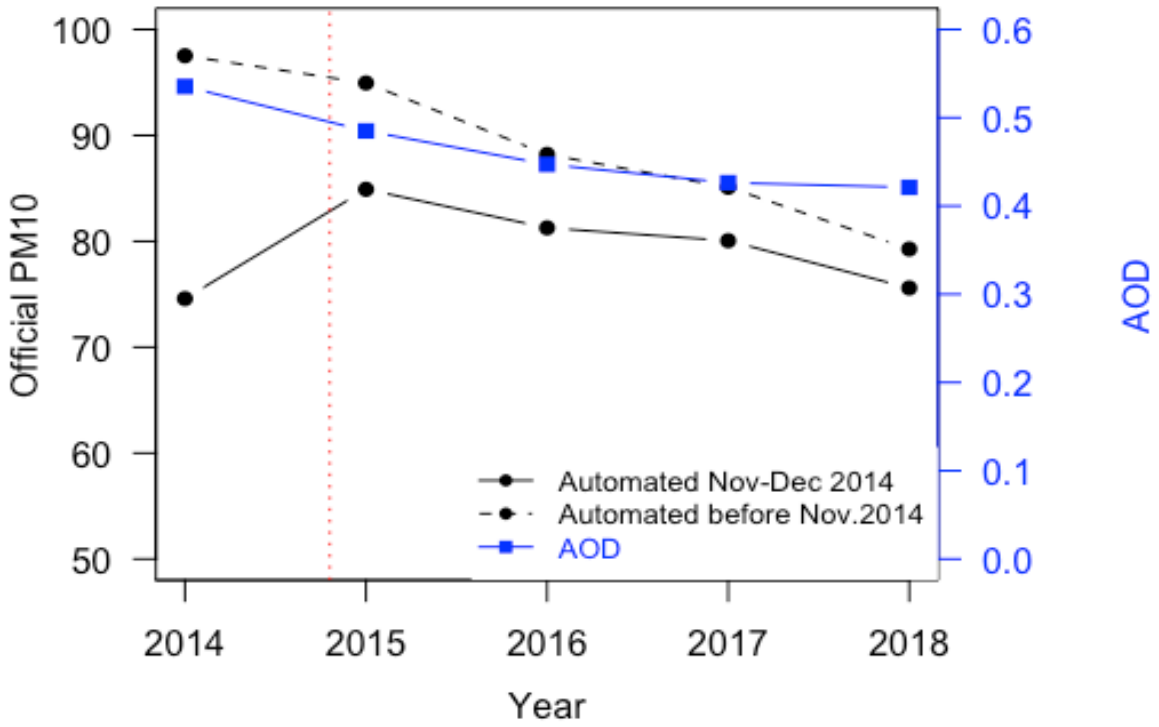


Figure 3-12 Annual PM_{10} and AOD from 2014 to 2018

To sum up, before automation, local environmental bureaus collected data and submitted it to the province and central authorities without validation. This created opportunities for local governments to manipulate air quality data, such as reporting a lower number than was accurate, especially at incentive cut-offs, resulting in a discontinuity at those cut-offs. In the new monitoring

system, opportunities for selective reporting are greatly mitigated as air quality data is sent to the central government in real-time., and it is thus less likely to manipulate the air pollution reports through soft misreporting. However, this advancement is not immune to *hard manipulation* that targets the monitoring facility before the data is collected and reported by the system. In the next section, I present and discuss the results with regard to the second type of possible manipulation, tampering with the station.

8.2 Does the reform reduce hard misreporting?

The second type of manipulation involves interfering with stations that collect air pollution data. Here, I estimate if the reform reduces hard misreporting using the DID SARAR (Spatial Auto-Regressive model with Auto-Regressive disturbances, or SARAR for short) estimator. The unit of analysis is station-month. The SARAR model includes spatially lagged air pollution as well as a spatially lagged error term to capture the spatial spillover and dependence of air pollution. I first conduct parallel trend test, and then I report the regression results from different specifications.

Similar to what was mentioned in the previous section, the analysis of the reform effect on reducing misreporting assumes that the growth rate of the PM_{10} in the absence of the centralization reform in treated stations would not differ from trends in non-treated stations. Violation of the parallel-trends assumption introduces a bias in difference-in-difference estimates of the treatment effect. While I cannot observe the growth rate of the PM_{10} in the absence of the reform for the treatment provinces, I can compare treated and control stations in the periods prior to the reform implementation. Figure 3-13 shows the predicted PM_{10} growth rate conditional on control variables such as AOD and weather conditions, using only the pre-period data and evaluating separately for treated and control stations. I find that the control station and treated station have a similar pre-trend in predicted PM_{10} growth before the reform. The treated station (red line) tends to have a lower growth rate, but the general pattern is similar to both the treated and control stations. This justifies, at least visually, the assumption of using difference in difference estimator.

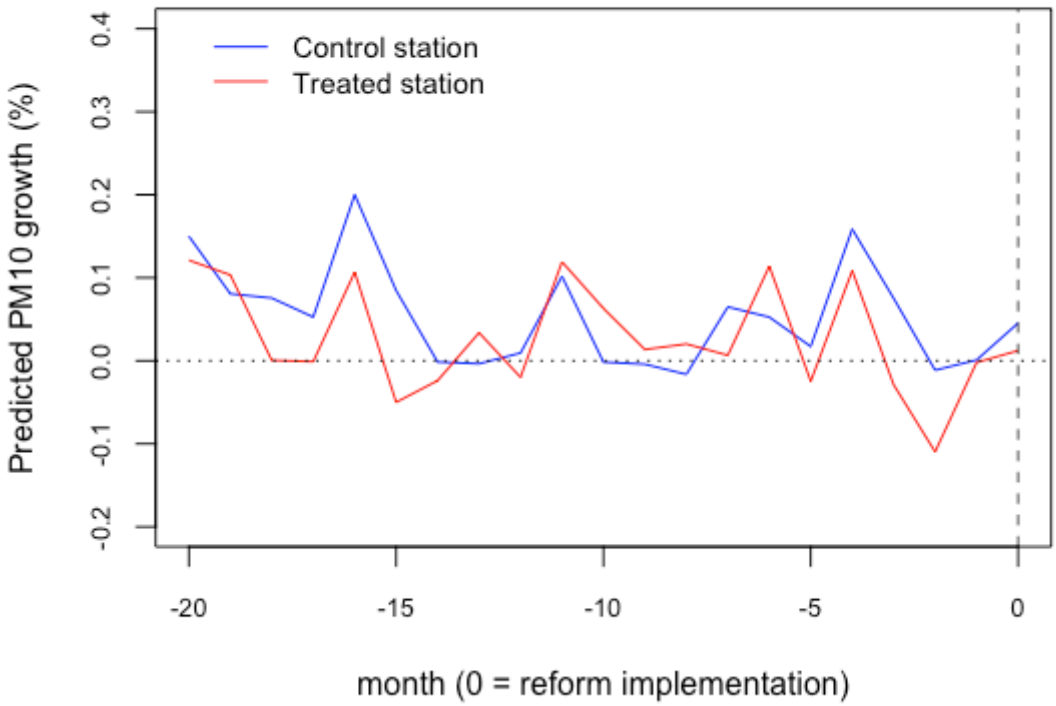


Figure 3-13 Parallel trend test (hard misreporting)

Table 3-2 shows the spatial panel regression results with different specifications. The estimator is spatial panel regression with spatially autocorrelated dependent variable I find that, on average, the treated station is associated with lower translation between AOD and PM_{10}, while the post-treatment period, on average, is associated with higher translation between AOD and PM_{10}, suggesting an overall improvement in pollution monitoring accuracy after the reform. In addition, the interactions with weather variables are largely consistent with aerodynamic literature. For example, higher air temperatures increase the ability to map from AOD to PM_{10}, while dew point temperature, which is related to humidity, and decreases such mapping ability. Furthermore, precipitation increases the mapping, whereas sky coverage significantly decreases such mapping.

Table 3-2 Spatial panel SARAR regression results

Dependent variable: PM_{10} growth rate	*Model 1*	*Model 2*	*Model 3*	*Model 4*
AOD growth rate	0.006***	0.005***	0.033**	0.035**
	(0.0007)	(0.001)	(0.025)	(0.025)
AOD growth rate × Post period × Treated	**0.005***	**0.004***	**0.006****	**0.009*****
	(0.002)	**(0.002)**	**(0.002)**	**(0.001)**
Post period × Treated	-0.001	-0.002	-0.002*	0.038***
	(0.003)	(0.003)	(0.003)	(0.005)
AOD growth rate × Treated	-0.002	-0.003*	-0.003	-0.004*
	(0.002)	(0.002)	(0.003)	(0.002)
AOD growth rate × Post period	0.006	0.006	0.007***	0.007**
	(0.009)	(0.009)	(0.001)	(0.003)
AOD growth rate ×Air Temperature			0.00003***	0.00003***
			(0.00002)	(0.00003)
AOD growth rate × *Dew* Temperature			-0.00003***	-0.00003**
			(0.00002)	(0.00003)
AOD growth rate × Windspeed			-0.00008	-0.0001
			(0.0001)	(0.0001)
AOD growth rate × 6 Hr. Precipitation			0.00004**	0.0001*
			(0.00001)	(0.0001)
AOD growth rate × Sky Coverage			-0.003***	-0.003**
			(0.001)	(0.001)
AOD growth rate × Pressure			-0.00003	-0.001*
			(0.00004)	(0.00003)
No. of observation (station - month)	47,904	47,904	47,904	47,904
Rho (spatial lag parameter)	0.83	0.82	0.83	0.75
Lambda (spatial error parameter)	/	0.19	/	0.20
Weather variables	Yes	Yes	Yes	Yes
Two-way FE	Yes	Yes	Yes	Yes
With spatial lag	Yes	Yes	Yes	Yes
With spatial error	No	Yes	No	Yes

Note: Dependent variable: monthly PM_{10} growth rate. Clustered standard error in parentheses. Significance level: *** p<0.01, ** p<0.05, * p<0.1.

We are primarily concerned with the effect of the reform on misreporting, which is captured by the coefficient on the interaction of *AOD growth rate, Post period*, and *Treated.* This variable is interpreted as how the difference in the ability to map from AOD to PM_{10} between treated units and control units differ after the treatment is implemented. This estimate is significant across specifications, yet the value itself does not have a direct interpretation due to the spatial dependence process: depending on each station's location in the system, each station will face different effects, and the average total spatial effect across all locations is more relevant. Based on the estimated coefficient β, the total spatial effect can be recovered by applying $(\mathrm{I} - \rho\mathrm{W})^{-1}\beta_k$ to the coefficient of interest. Table 3-3 shows the spatial direct effect (the average effect on the PM_{1o} after the post-period if the unit being evaluated is treated, or the average of the diagonal elements of the matrix $(\mathrm{I} - \rho\mathrm{W})^{-1}\beta_k$, indirect effect (the average effect on PM_{10} pollution if the neighbor of the unit being evaluated is treated, or the average of the off-diagonal matrix), and the total effect (the sum of direct and indirect effects, or the average of all elements of the matrix). The coefficients are shown in percentage format because both AOD and PM_{10} enter the equation as a growth rate.

Table 3-3 Spatial effect (dependent variable: monthly PM_{10} growth rate)

	AOD growth rate × Post period × Treated		
	Direct Effect	*Indirect Effect*	*Total Effect*
Model 1	0.61%	0.32%	0.93%
Model 2	0.23%	0.27%	0.50%
Model 3	0.97%	0.50%	1.47%
Model 4	1.24%	1.34%	2.59%

On average, after the reform implementation, a one percent increase in the actual pollution (AOD) at the treated stations leads to higher reported pollution PM_{10}, around 1.3% higher, suggesting a more truthful reporting of the actual pollution for the treated station after the reform compared to the control station. This effect is significant across specifications, and with spatially correlated error terms. As a robustness check, I also repeat the above model with a distance-based weight matrix using 5 km as the critical band. The results are reported in the Appendix and the findings remain largely consistent.

Figure 3-14 shows that stations that participated in the reform experienced a substantial increase in the ability to translate AOD into reported pollution in the post-treatment period as

characterized by the positive treatment effect. Such treatment effects are significantly different from zero in the post-treatment period.

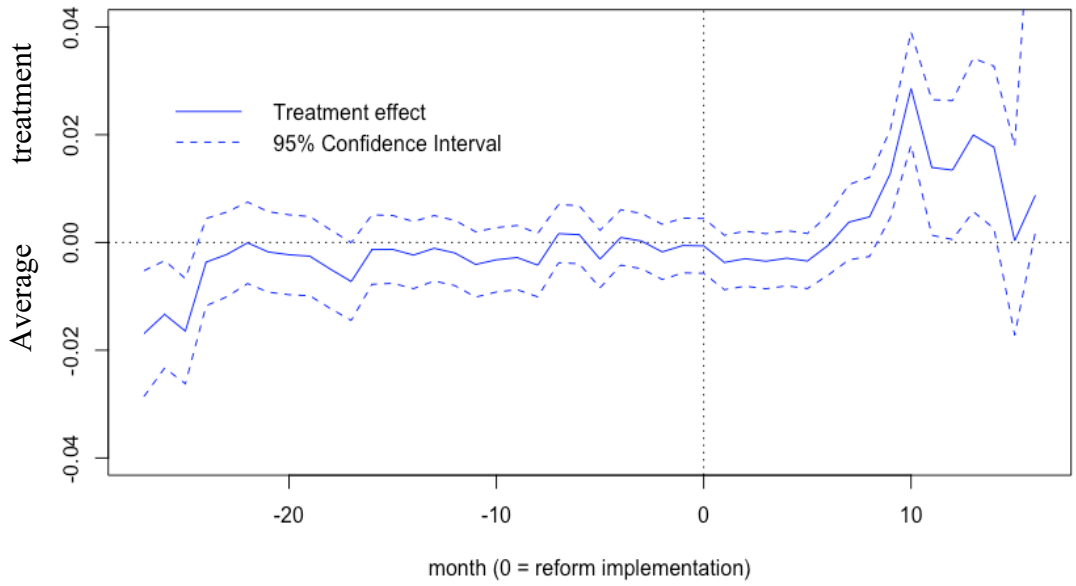


Figure 3-14 Treatment effect by month (hard misreporting, 5 nearest neighbors)

I also noticed some treatment effect delays after the reform was implemented. The significant treatment effect did not appear until approximately 5 months after the treatment month. Starting from the ninth month after the implementation of the reform. After the reform, the same amount of increase in actual pollution now maps into higher reported pollution in treated stations, implying that the same amount of increase in actual pollution now maps into higher reported pollution in treated stations. Such an effect peaked in the tenth month following the reform, then declined in 2018, but the effect remains significant. This is likely because institutional reform and the restructuring of the environmental bureaucracy take time and may cause a delay in reaching effect.

The results above demonstrate the relationship between true pollution and reported pollution at the mean of their distribution. A related type of hard misreporting, as suggested by the news report and anecdotal evidence, happens at the right tail of polluted days or days with heavy pollutions. Several cities have reported having interference with the station during heavily polluted days. This suggests manipulation that happens at the tail of the pollution distribution. To quantify manipulation during days with severe pollution, I adopt spatial quantile regression at the 75%, 85%, and 95% quantiles of the PM_{10}. I use Kim and Muller's Two-Stage Quantile estimator to

account for the spatial lagged term of the AOD. Figure 3-15 below shows the results. First, I observe that, overall, there is an increase in the ability to predict PM_{10} from AOD at the tail, suggesting a gradual improvement in the accuracy and reliability of the reported data during heavily polluted days, yet such an overall trend is not statistically significant.

Second, at different quantiles over the PM_{10} distribution, the results show that, on average, one unit increase in true pollution (AOD) translates into more reported pollution (PM_{10}. In other words, the estimated misreporting through hard misreporting during the heavily polluted days reduced by 0.05 units in the PM_{10} at the 75% quantile (orange line, top left plot), 1.9 units at the 85% quantile (green line, top right plot), and 5.9 at the 95% quantile (purple line, bottom left plot). In all three plots, the bootstrapped standard error gives the 95% confidence interval (marked by the dashed lines). Although the average for any quantitative is not significant, the effect on the 95% eight months after the reform's implementation is significant at the 5% level. Notice that the largest (and the most immediate) effect seems to happen during the more polluted days, or the 95% quantile of the pollution. This suggests that the local government tends to target at the heavily polluted days in order to improve the air quality statistics.

To summarize, the results of the spatial quantile regressions indicate that the reform has a limited impact on reducing hard misreporting during heavy pollution days, but it does appear to reduce misreporting on extremely polluted days at the 95% tail of the PM10 distribution several months after implementation.

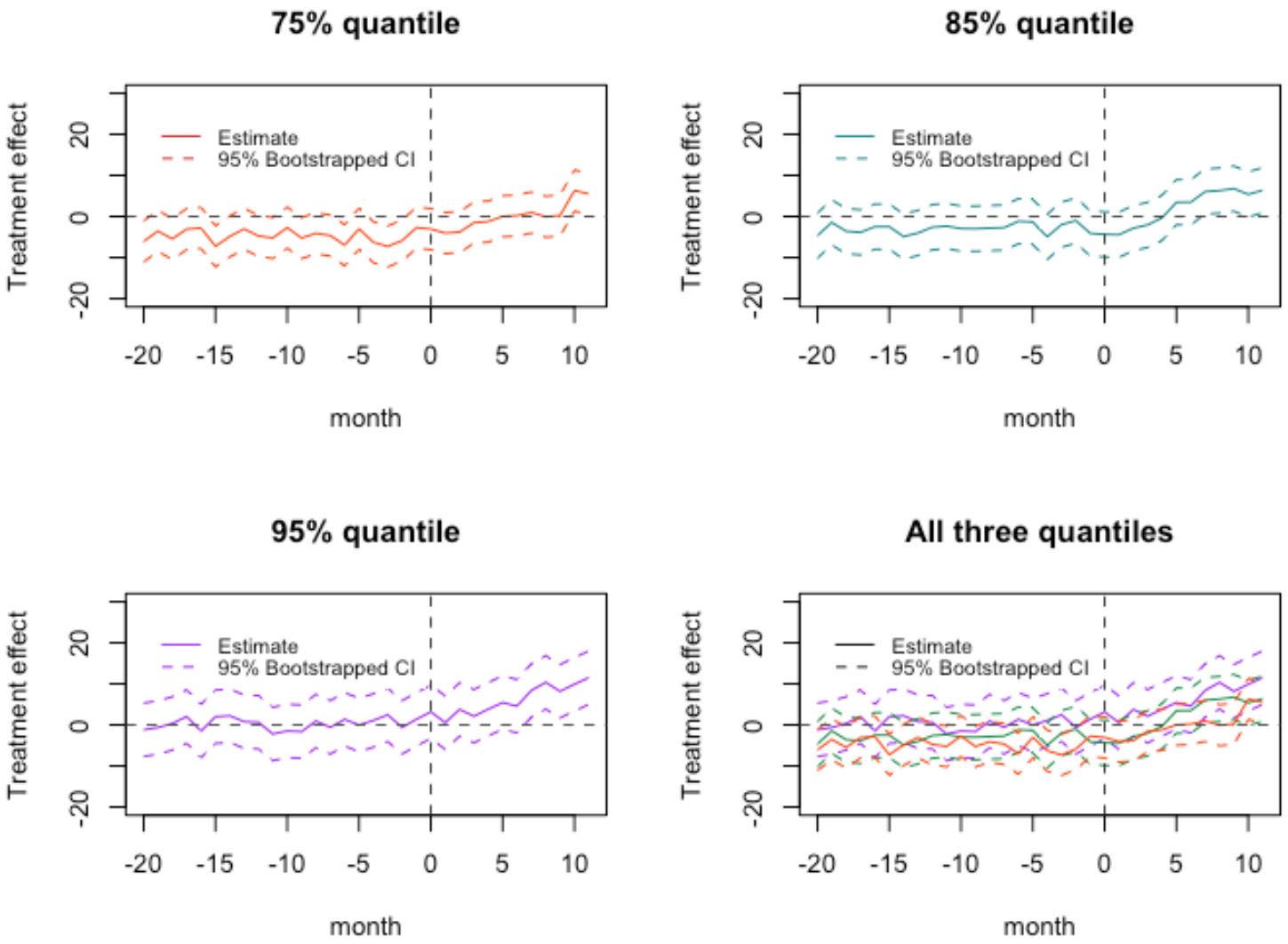


Figure 3-15 Treatment effect by month (Hard misreporting at the tail of the pollution)

8.2.1 Placebo test

For the placebo test, I first drop all the outcomes for treated observations after they receive treatment. Everyone in the remaining data only have untreated outcome data. Then I insert multiple phantom treatment events in multiple time periods of the remaining data for the treated observations. I run the same diff-in-diff model as before and I check the interaction coefficients. Figure 3-16 shows the placebo test. In the months prior to the reform, I did not observe significant reform effect. This lends credence to the difference in difference estimator.

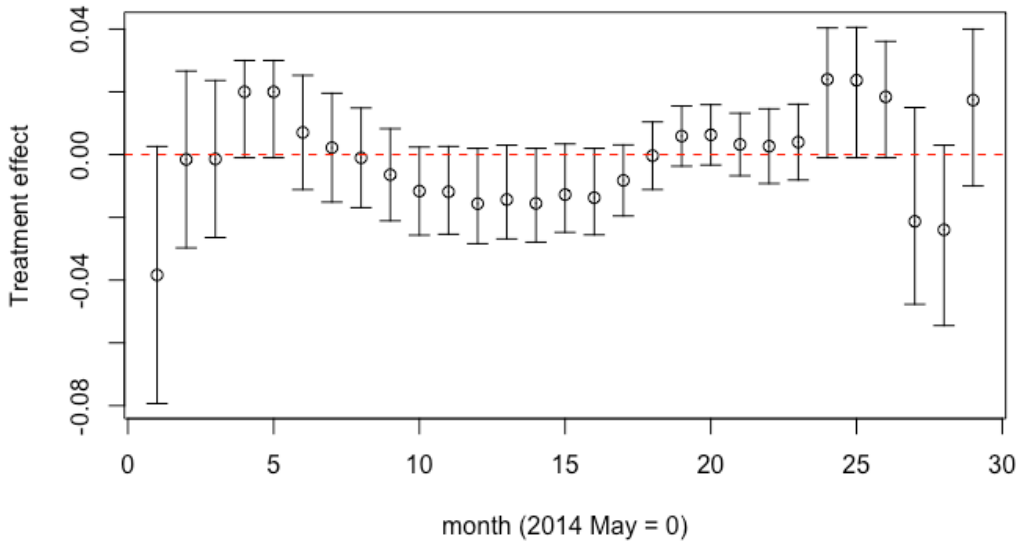


Figure 3-16 Placebo test

8.3 Robustness check

8.3.1 Selection bias

It is possible that the treated provinces self-sorted in the reform. For example, regions that would benefit more from participation (such as those who are nominated for "Environmental Protection Model Cities") might be more likely to join the reform. This is a concern about the validity of the parallel trend assumption. If such drivers also correlate with their misreporting behavior, it may bias the treatment effect estimation. It is important, therefore, to construct a credible control group so that outcomes can be reasonably compared. To reduce the impact of self-selection, I incorporate propensity score matching before estimating the Difference-in-Difference model.

I use propensity score to improve the balance of the treated and control regions. The propensity score is the probability of treatment assignment conditional on observed baseline characteristics. The propensity score allows us to further analyze this observational setting so that it mimics some of the characteristics of a randomized treatment assignment conditional on observables. I first estimate the propensity score using Logit regression for each province. I use GDP growth, GDP per capita, PM10 level, AOD level, GPD composition, number of workers employed in the different sectors, population density, and population as predictors. All variables are measured in the pre-treatment periods. The results are shown in Figure 3-17.

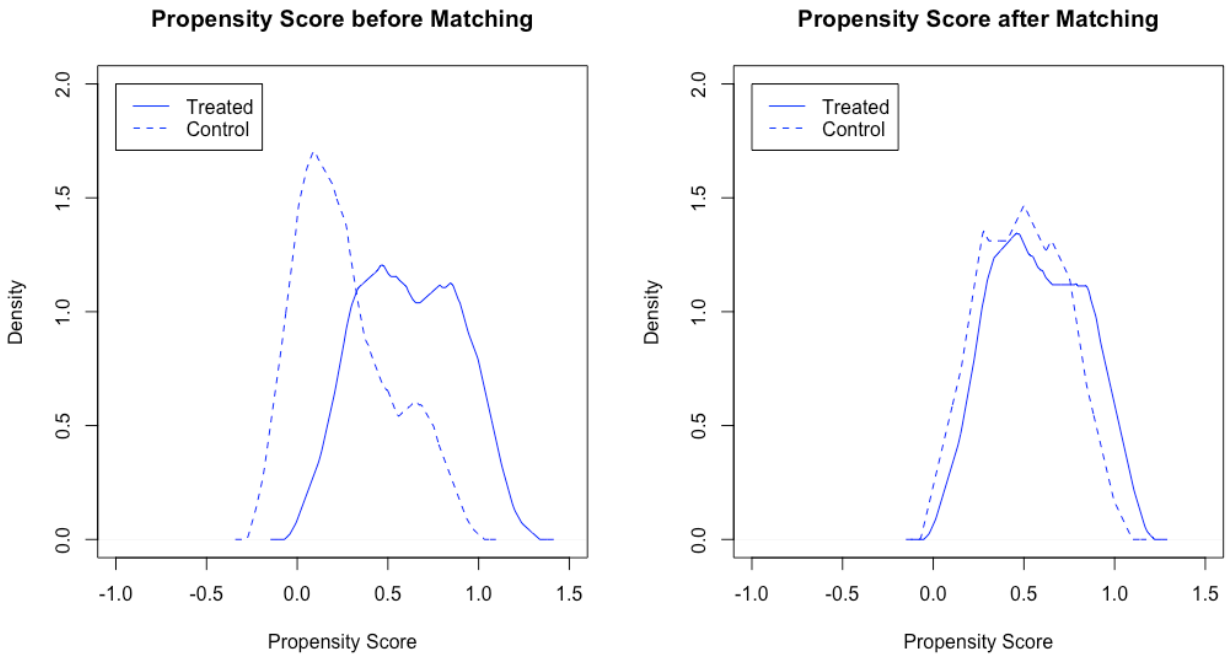


Figure 3-17 Propensity Score before and after the matching
(Left panel: before the matching. Right panel: after the matching)

The left figure in Figure 17 shows the density of propensity scores before matching. There is an evident imbalance between the treated and the control observations. Then I implement caliper matching with replacement using the fitted probability from the first stage. The right figure in Figure 13 shows the density of the score after matching. There is a significant improvement in balance after matching, suggesting very similar control and treated provinces after accounting for selection on observables. The test for equality of densities as described in Li, Maasoumi, and Racine (2009) shows that these two densities are significantly different at the 5% level.

I then use the estimated propensity score in the nearest neighbor matching to create similar control and treated groups and estimate the average treatment effect between the matched provinces. Recall that one-third of the provinces were treated, and two-thirds were control provinces. Following the matching process, there were a total of 16 matched provinces, with 8 controls and 8 treated provinces. Among the matched provinces, Hebei, Shanghai, Jiangxi, Shandong, Hubei, Guangdong, Chongqing, and Shaanxi are treated, while Inner Mongolia, Zhejiang, Henan, Hunan, Tibet, Ningxia, Guizhou, and Xinjiang are controlled.

Table 3-4 below compares the key characteristics of the matched and unmatched provinces. First, notice that matched control provinces tend to have a higher PM_{10} than the unmatched control provinces (the p value = 0.02), as the treated provinces tend to have a higher PM_{10} than the control provinces. Moreover, the matched control provinces tend to have a higher GDP growth rate than the unmatched control provinces (the p value = 0.01), as the treated provinces tend to have a higher GDP growth rate than the control provinces on average. With regards to the location of the provinces, the map shows the matched and unmatched provinces, both treated and control provinces. Other than these covariates, the rest of the covariates have also been better balanced. Overall, after the propensity score matching, the left sample represents a group of faster growing provinces with higher GDP per capita, with higher pollution (although not the case for AOD), much lower agricultural employment and lower GDP coming from agricultural sector, and more manufacturing firms. Some of the largest provinces in the north and northwest, such as Xinjiang, Tibet, and Inner Mongolia, are also among the matched control provinces. Figure 3-18 shows the location of the matched and unmatched provinces.

Table 3-4 Descriptive statistics by matched and unmatched provinces

	Treated		Control	
	Matched	*Unmatched*	*Matched*	*Unmatched*
AOD	137.9	123.2	102.9	107.0
PM_{10}	90.9	88.6	92.6	74.4
Air temperature (Celsius)	165.5	153.7	156.2	158.2
Wind Speed (km/hour)	25.9	26.6	26.0	27.5
Precipitation (mm)	34.8	32.1	31.4	33.6
GDP (10 thousand Yuan)	¥60,026,571	¥45,958,040	¥6,249,054	¥6,623,496
GDP growth rate (%)	8.6%	9.7%	8.4%	6.2%
Percentage of agricultural sector in GDP (%)	4.9%	2.8%	4.7%	8.3%
Percentage of Mfg sector in GDP (%)	46.9%	46.1%	43.2%	43.1%
Percentage of service sector in GDP (%)	48.1%	51.1%	52.1%	48.5%
Per capita GDP (Yuan)	¥70,619	¥88,842	¥76,630	¥52,903
Total retail sale (10 thousand Yuan)	¥24,595,437	¥15,352,819	¥3,190,739	¥3,711,245
Number of Mfg firms (number)	2267.2	1911.5	352.2	206.8
Total product of Mfg firms (10 thousand Yuan)	¥86,344,278	¥78,121,221	¥7,298,629	¥7,391,668
% Employment of agricultural sector (%)	0.4	0.3	1.1	2.7
% Employment of Mfg sector (%)	46.9	50.9	37.6	40.7
% Employment of service sector (%)	52.7	48.7	61.6	56.8
Fiscal income (10 thousand Yuan)	¥9,360,033	¥4,305,896	¥624,462	¥593,669
Fiscal expenditure (10 thousand Yuan)	¥11,323,583	¥5,437,733	¥861,998	¥1,044,468

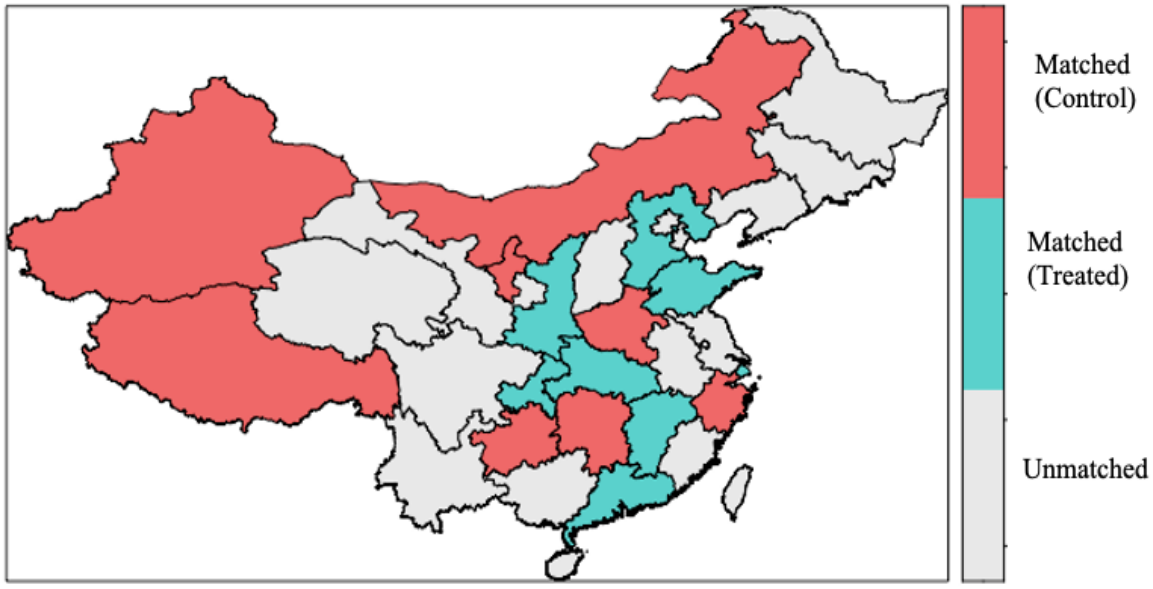


Figure 3-18 Location of the matched and unmatched provinces.

Figure 3-19 shows the treatment effects across months using only matched provinces. The magnitude of the treatment effect is greater than the unmatched model. I also observe a much earlier treatment effect that occurred right after the implementation of the reform. Recall that in the unmatched model, the treatment effects are observed 9 months after the reform was implemented.

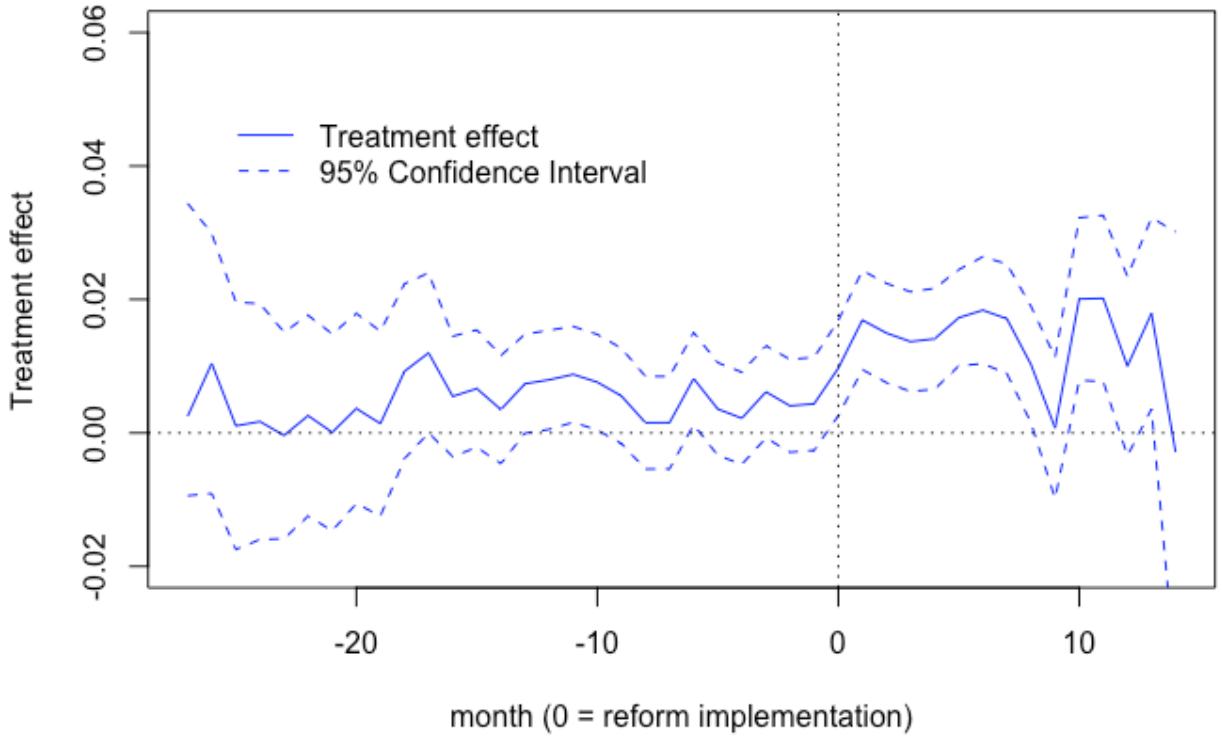


Figure 3-19 Treatment effect by month (Hard misreporting, matched

8.3.2 SUTVA discussion

One assumption of conducting a valid DID analysis is the Stable Unit Treatment Values Assumption or SUTVA. SUTVA requires that the response of a particular observation depends only on the treatment to which it was assigned, not the treatments of others. In our context, there is a risk that the SUTVA is violated because untreated provinces can be affected by the reform implemented in other provinces. For example, with the limited investigation and auditing budget, the central government might choose to audit the untreated cities. Therefore, those cities that did not participate in the reform might then change their misreporting behavior. Such a potential spillover effect stems from general equilibrium and, consequently, I might observe a reduction in misreporting for both treated and control cities.

To account for the possibility of SUTVA, I assume that once the reform is implemented, all the provinces and their stations, regardless of whether they are participating or not, have been affected. Table 3-5 shows the average effect of treatment after the reform (one difference) assuming all the stations are treated. I do not find a significant post-period effect, and the total effect is much smaller, suggesting that the spillover effect, if any, is very limited.

Table 3-5 Spatial effect (assuming all stations are treated after post-period)

	AOD growth rate × Post period = 1		
	Direct Effect	*Indirect Effect*	*Total Effect*
Model 1	0.12%	0.20%	0.32%
Model 2	0.12%	0.32%	0.44%
Model 3	0.10%	0.18%	0.28%
Model 4	0.10%	0.18%	0.28%

8.3.3 Individual treatment timings

I further extend the speciation in the equation above to a Difference-in-differences with variation in treatment timing. This is not considered the main model because (1) in China, the release of a national guideline is usually considered the official start of a reform, whereas the different timings set by the local government are to provide more implementation details, (2) the definitions of average treatment effects are more complicated in this case because the counterfactual is more complicated, and (3), the results show that the treatment effect based on individual treatment timings give much smaller estimated effect, consistent with the second point and thus this section is not the preferred model specification.

Intuitively, with different treatment timings, if a unit is treated very early or very late within the time frame, it will be given a smaller weight as treated but larger a weight as a control. If a unit is treated at some point of time in the middle of the panel, it will be given a larger weight. This can be regarded as a DID model with a heterogeneous treatment effect. With this set-up, the "control" monitors are never treated during this period and provide a plausibly credible counterfactual for the "treatment" monitors. Further, this is one approach to confronting the challenges associated with the staggered assignment of treatment. Two-way fixed effects DID model with varying treatment timings across treated units, which essentially is a weighted average of all possible standard DD estimators that compare different timing groups to each other. When

treated observations experience treatment at different times, one can not estimate a standard DID because the post-treatment period dummy can not be defined for control observations.

A standard DID estimate is the difference between the change in outcomes before and after treatment (first difference) in treatment versus the control group (second difference), under a common trends assumption, a two-group/two-period (2x2) DID identify the average treatment effect on the treated, while a generalized DID like the one in our paper is an extension of standard DID that allows different treatment timings. A two-way fixed effect DID is the weighted average of all possible combinations of treated and control groups at different timings with one another. Some use units treated at a particular time as the treatment group and untreated units as the control group. Some compare units treated at two different times, using the later-treated group as a control before its treatment begins and then the earlier group as control after its treatment begins.

Table 3-6 displays the results of using a single city's treatment timings. I did not find a significant impact of individual treatment timing on the ability to translate AOD into PM_{10}, suggesting that it is the national guideline release (i.e., July 2016) that marks when the impact of the reform starts.

Table 3-6 Spatial panel SARAR results with province-specific reform timing

Dependent variable: PM_{10} growth rate	*Model 1*	*Model 2*	*Model 3*	*Model 4*
AOD growth rate	0.002***	0.002***	0.573*	0.913**
	(0.0004)	(0.001)	(0.253)	(0.304)
AOD growth rate × Post period × Treated	**-0.001**	**0.001**	**-0.001**	**0.001**
	(0.002)	**(0.002)**	**(0.002)**	**(0.002)**
Post period × Treated	0.015**	0.02***	0.013**	0.015**
	(0.005)	(0.005)	(0.005)	(0.005)
AOD growth rate × Treated	-0.004	-0.013*	-0.005	-0.015*
	(0.003)	(0.007)	(0.003)	(0.007)
AOD growth rate × Post period	0.002	0.001	0.002*	0.002
	(0.001)	(0.002)	(0.001)	(0.002)
AOD growth rate ×Air Temperature			0.00001	0.0001*
			(0.00002)	(0.00003)
AOD growth rate × *Dew* Temperature			-0.00002	-0.0001*
			(0.00002)	(0.00003)
AOD growth rate × Windspeed			0.00002	-0.00002
			(0.0001)	(0.0001)
AOD growth rate × 6 Hr. Precipitation			-0.00004	-0.0001
			(0.0001)	(0.0001)
AOD growth rate × Sky Coverage			-0.002**	-0.002*
			(0.001)	(0.001)
AOD growth rate × Pressure			-0.0001*	-0.0001***
			(0.00002)	(0.00003)
No. of observation (station - month)	47,904	47,904	47,904	47,904
Rho (spatial lag parameter)	0.69	0.68	0.69	0.68
Lambda (spatial error parameter)	/	0.01	/	0.2
Weather variables	Yes	Yes	Yes	Yes
Two-way FE	Yes	Yes	Yes	Yes
With spatial lag	Yes	Yes	Yes	Yes
With spatial error	No	Yes	No	Yes

Note: Dependent variable: monthly PM10 growth rate. Clustered standard error in parentheses. Significance level: *** p<0.01, ** p<0.05, * p<0.1

9 Policy implications

The empirical results from this paper have several important implications for reducing government misreporting. First, in terms of reform effect, the findings indicate that this environmental centralization reform is likely to have successfully reduced hard misreporting, both on regular and heavily polluted days. Before the reform, evidence shows that many monitoring facilities are subject to human intervention, for example, through covering the pollution sampling vents, spraying water or chemical solution on the facility, etc., in order to reduce the pollution readings. Based on difference in different model with spatially lagged pollution, the results show that the hard misreporting has reduced (more precisely, the ability to map the reported pollution from true pollution has significantly improved for treated stations after the reform). This suggests that the hard misreporting caused by tampering with the station and the surrounding environment has been reduced.

On the other hand, there is no evidence of the effect of the reform on reducing soft misreporting. This is likely due to the fact that the installation of automatic monitoring equipment prior to the reform. The improvement in surveillance technology allows better control of data measurement and transmission, as well as quality assurance and quality control, and has increased the technical difficulty of misreporting through altering the readings. In addition, there is significant evidence from spatial quantile regressions that suggests less misreporting during heavily polluted days. Overall, the results imply that the reform has reduced one major source of biased pollution statistics both during regular days and heavy polluted days.

From a policy standpoint, the findings show that reinforcing control over local agents, preferably in conjunction with improved technology, is an effective approach that appears to be capable of mitigating misreporting at the local level. In particular, the distortive effect of performance-based evaluation can be partially mitigated through institutional restructuring that centralizes power away from the agent. This can be used to better align objectives between the principal and the agent, preventing the scope of the opportunistic behavior of the agent and achieving better performance.

This paper suggests that, through certain authority arrangements in the system, misreporting can be avoided from the outset. The authority arrangement is more appealing than other approaches for preventing fraudulent reporting and data distortion because it is less expensive and less likely to produce unintended consequences. This implication is consistent with

the scholarly literature on economics. Studies have considered centralization or decentralization as an important ex-ante control that can be used to induce desirable behavior of political agents (Moe, 2013). Specifically, when compared to other approaches, the authority arrangement is more appealing for preventing distortion because it is less expensive and less likely to produce unintended consequences. Auditing or punishment, for example, which is an ex-post approach to correcting distortion or dishonesty, is associated with high enforcement costs; weaker political awards can reduce the potential benefit of distortion but can also result in flat incentive that leads to passive agents. Therefore, centralization/decentralization stands out from other institutional approaches in reducing the potential distortive behavior of agencies.

10 Conclusion and future studies

This paper empirically examines if centralization affects local misreporting. There has been an emerging literature on government misreporting, including Martinez (2018), Edmond & Lu (2018), Kalgin (2016), Ghanem & Zhang (2014). The findings in this paper shed light on how to improve government accountability through proper power distribution. There have been numerous efforts around the world to reduce government misreporting in order to promote transparency and accountability. Most of the initiatives and programs have focused on improving public access to government data. Some are designed to preventing misinformation, misreporting, and misrepresentation. Literature on how authority distribution may affect local agencies behavior, such as Oates (1999), Besley & Prat (2006), and Evdokimov & Garfagnini (2018). The logic is that centralization, either fiscal or administrative or both, reinforces the control of upper-level authority over its local bureaucracies, limiting the scope for opportunistic behavior of the latter (Mertha, 2005) such as misreporting. Yet this speculation has not been empirically examined.

This paper studies this question in the context of a recent environmental reform in China. This environmental centralization reform is also known as vertical environmental reform. In China, the air pollution information is primarily monitored and collected by city environmental protection bureaus. Before the reform, these bureaus are administered by city governments that are rewarded for meeting pollution abatement targets. This might incentivize the city government to skew air pollution statistics, making air pollution vulnerable to manipulation. After the reform, the city environmental bureau is no longer de jure controlled by the city government and should ideally have more independence in reporting air pollution truthfully.

To assess the impact of this reform, I first quantified air pollution data misreporting, then examined the treatment effect of the reform on that measure of manipulation. The measure of air pollution manipulation in this paper expands on previous studies by allowing for misrepresentation that may or may not create a discontinuity in the distribution of air pollution data. The reform, in my opinion, reduces "hard misreporting"--those who manipulate air pollution reports by interfering with the stations. This finding is robust to a battery of additional checks, such as accounting for selection bias, different spatial weight matrices, and different reforms for individual provinces. On the other hand, there is no evidence of the effect of the reform on reducing soft misreporting that modifies air pollution data after it is collected. This is likely due to the fact that the installation of automatic monitoring equipment prior to the reform has increased the technical difficulty of misreporting through altering the readings.

One limitation of this study is the unbalanced data from the pollution monitoring stations. Since more than half of the current stations were installed gradually. The locations of the newly established stations are unlikely to have been randomly chosen and might be related to other factors that affect pollution levels. Therefore, to maintain a comparable sample, I removed all the stations that are installed after 2014. This results in the removal of more than 700 stations. For future studies, it would be helpful to account for such non-random missing stations and use the additional information from the newly installed stations to further test the robustness of the results. In addition, the spatial weight matrix does not account for the wind direction, although this is unlikely to bias our results, it would be useful to allow asymmetric spatial neighbor that account for different wind direction to reduce the variance of our estimation.

Also, the reform probably did not provide the central government with full control over the local environmental collection, as I discovered a discontinuity after the reform, implying that the central government does not have complete control over the data. Also, the results show that although hard misreporting decreases, it did not disappear after the reform. This further indicates that the reform raises the cost of misreporting but does not make it impossible to fully eliminate misreporting, and that the central government still does not have full and direct control over pollution data measurement.

Finally, it is also possible that after the reform, the rules of the reporting game have changed, and there might be new types of manipulation that our empirical estimation is not designed to capture. Anecdotal evidence suggests that recent manipulation has targeted software

automation. Moreover, it is possible that the reform will induce strategic coordination among cities if the cities all respond strategically to the reform. Unfortunately, this is beyond the scope of this study and our model is not designed to model such strategic coordination among local governments. Future research on this topic would benefit from more data and a deeper understanding of the incentive structure, as well as new measures of manipulation.

11 References

A straightforward case of fake statistics. (2019, April 18). Retrieved from Review of African Political Economy: http://roape.net/2019/04/18/a-straightforward-case-of-fake-statistics/

Alonso, R., Dessein, W., & Matouschek, N. (2008). When does coordination require centralization? *American Economic Review, 98*(1), 145-179.

Andrews, S. (2008). Inconsistencies in air quality metrics: 'Blue Sky'days and PM10 concentrations in Beijing. *Environmental Research Letters, 3*(3).

Angelucci, M., & Di Maro, V. (2016). Programme evaluation and spillover effects. *Journal of Development Effectiveness, 8*(1), 22-43.

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy, 100*(3), 598-614.

Balafoutas, L., Dutcher, E. G., Lindner, F., & Ryvkin, D. (2017). The optimal allocation of prizes in tournaments of heterogeneous agents. *Economic Inquiry, 55*(1), 461-478.

Batterbury, S., & Fernando, J. (2006). Rescaling governance and the impacts of political and environmental decentralization: an introduction. *World development, 34*(11), 1851-1863.

Baye, M. R., Kovenock, D., & De Vries, C. G. (1993). Rigging the lobbying process: an application of the all-pay auction. *The American Economic Review, 83*(1), 289-294.

BBC. (2018, August 29). *Puerto Rico increases Hurricane Maria death toll to 2,975*. Retrieved July 10, 2019, from BBC: https://www.bbc.com/news/world-us-canada-45338080

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551-572.

Besley, T., & Prat, A. (2006). Handcuffs for the grabbing hand? Media capture and government accountability. *American economic review, 96*(3), 720-736.

Blonz, J. A. (2019). The Welfare Costs of Misaligned Incentives: Energy Inefficiency and the Principal-Agent Problem. *Working paper*.

Bohte, J., & Meier, K. J. (2000). Goal displacement: Assessing the motivation for organizational cheating. *Public Administration Review, 60*(2), 173-182.

Brombal, D. (2017). Accuracy of environmental monitoring in China: exploring the influence of institutional, political and ideological factors. *Sustainability, 9*(3), 324.

Brombal, D. (2017). Accuracy of environmental monitoring in China: Exploring the influence of institutional, political and ideological factors. *Sustainability, 9*(3), 324.

Carassava, A. (2010, September 23). *Greece condemned for falsifying data*. Retrieved from Financial Time: https://www.ft.com/content/33b0a48c-ff7e-11de-8f53-00144feabdc0

Dessein, W. (2002). Authority and communication in organizations. *The Review of Economic Studies, 69*(4), 811-838.

Eaton, S., & Kostka, G. (2014). Authoritarian Environmentalism Undermined? Local Leaders' Time Horizons and Environmental Policy Implementation in China . *China's Quaterly*, 359-380.

European Council. (2015, July 13). *Deficit data in Valencia: Spain fined for misreporting*. Retrieved from European Council: https://www.consilium.europa.eu/en/press/press-releases/2015/07/13/deficit-data-valencia/

European Council. (2018, May 28). *Austria fined for misreporting government debt data*. Retrieved from European Council: https://www.consilium.europa.eu/en/press/press-releases/2018/05/28/land-salzburg-austria-fined-for-misreporting-government-debt-data/

Fisman, R., & Gatti, R. (2002). Decentralization and corruption: evidence across countries. *Journal of Public Economics, 83*(3), 325-345.

Fisman, R., & Wang, Y. (2017). The distortionary effects of incentives in government: Evidence from China's" death ceiling" program. *American Economic Journal: Applied Economics, 9*(2), 202-18.

Friesen, L. (2003). Targeting enforcement to improve compliance with environmental regulations. *Journal of Environmental Economics and Management, 46*(1), 72-85.

GAO. (2019). *2019 High Risk List*. Retrieved June 19, 2019, from U.S Government Accountability Office: https://www.gao.gov/highrisk/overview

Ghanem, D., & Zhang, J. (2014). Effortless perfection: Do Chinese cities manipulate air pollution data? *Journal of Environmental Economics and Management, 68*(2), 203-225.

Ghanem, D., Shen, S., & Zhang, J. (2020). A Censored Maximum Likelihood Approach to Quantifying Manipulation in China's Air Pollution Data. *Journal of the Association of Environmental and Resource Economists, 7*(5), 965-1003.

Hölmstrom, B. (1979). Moral hazard and observability. *The Bell journal of economics*, 74-91.

Kalgin, A. (2016). Implementation of performance management in regional government in Russia: evidence of data manipulation. *Public Management Review, 18*(1), 110-138.

Kräkel, M., & Schöttner, A. (2010). Technology choice, relative performance pay, and worker heterogeneity. *Journal of Economic Behavior & Organization, 76*(3), 748-758.

Kumar, M. (2019, May 9). *India's incredulous data: Economists create own benchmarks*. Retrieved from Reuters: https://in.reuters.com/article/us-india-economy-data-insight/indias-incredulous-data-economists-create-own-benchmarks-idINKCN1SF0L6

LeBeau, C. (2018, May 16). *Ricardo Perez-Truglia examines what happened when the Argentine government lied about inflation numbers*. Retrieved from UCLA Anderson Review: https://www.anderson.ucla.edu/faculty-and-research/anderson-review/inflation-lies

Lieberthal, K., & Oksenberg, M. (1990). *Policy making in China: Leaders, structures, and processes.* Princeton University Press.

Ma, B., & Zheng, X. (2018). Biased data revisions: Unintended consequences of China's energy-saving mandates. *China Economic Review, 48*, 102-113.

Ma, Y. (2017). Vertical environmental management: A panacea to the environmental enforcement gap in China? *Chinese Journal of Environmental Law, 1*(1), 37-68.

Macho-Stadler, I., & Perez-Castrillo, D. (2006). Optimal enforcement policy and firms' emissions and compliance with environmental taxes. *Journal of Environmental Economics and Management, 51*(1), 110-131.

Martinez, L. (2018). How much should we trust the dictator's GDP estimates? *Working Paper*.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2), 698-714.

McGee, R., & Gaventa, J. (2011). Shifting power? Assessing the impact of transparency and accountability initiatives. *IDS Working Papers,, 383*, 1-39.

Mertha, A. C. (2005). China's "soft" centralization: shifting tiao/kuai authority relations. *The China Quarterly, 184*, 791-810.

Ministry of Ecology and Environment of China. (2018, 3 28). *关于部分城市环境空气质量自动监测站点受到人为干扰有关情况的通报* . Retrieved from Ministry of Ecology and Environment of China: https://www.mee.gov.cn/gkml/sthjbgw/stbgth/201803/t20180329_433243.htm

Ministry of Ecology and Environment of China. (2018, 08 28). *Report on Linfen City, Shanxi Province misreporting air pollution data of national controlled air pollution monitoring records*. (E. M. Office, Producer) Retrieved from https://www.mee.gov.cn/xxgk2018/xxgk/xxgk06/201808/t20180830_629831.html

Miranda-Zanetti, M. D. (2019). Tampering with inflation data: A Benford law-based analysis of national statistics in Argentina. *Physica A: Statistical Mechanics and its Applications, 525*, 761-770.

Moldovanu, B., & Sela, A. (2006). Contest architecture. *Journal of Economic Theory, 126*(1), 70-96.

Mookherjee. (2015). Political Decentralization. *Annual Review of Economics*, 231-243.

Oates, E. W., & Schwab, R. M. (1988). Economic competition among jurisdictions: efficiency enhancing or distortion inducing? *Journal of public economics, 35*(3), 333-354.

Oates, W. (1999). An essay on fiscal federalism. *Journal of Economic Literature, 37*(3), 1120-1149.

People Daily. (2018, 12 1). *空气质量监测怎么干？监测项目更精细.* Retrieved from xinhuanet: http://www.xinhuanet.com/2018-12/01/c_1123792489.htm

Ran, R. (2017). Understanding blame politics in China's decentralized system of environmental governance: actors, strategies and context. *The China Quarterly, 231*, 634-661.

Reuters. (2020, November 10). *Fact check: Deviation from Benford's Law does not prove election fraud*. Retrieved from Reuters: https://www.reuters.com/article/uk-factcheck-benford/fact-check-deviation-from-benfords-law-does-not-prove-election-fraud-idUSKBN27Q3AI

South China Morning Post. (2018, 10 22). *Chinese regions accused of faking efforts to curb environmental problems*. Retrieved from South China Morning Post: https://www.scmp.com/news/china/politics/article/2169643/chinese-regions-accused-faking-efforts-curb-environmental

Stoerk, T. (2016). Statistical corruption in Beijing's air quality data has likely ended in 2012. *Atmospheric Environment, 127*, 365-371.

United Nation. (2002). In V. Cistulli, *Environment in decentralized development: Economic and institutional issues* (p. Chapter 2). Food & Agriculture Org.

United Nation. (2015). *United Nations Fundamental Principles of Official Statistics.* United Nation.

United Nations. (2014). Fundamental principles of official statistics. United Nations General Assembly.

Weaver, R. K. (1986). The politics of blame avoidance. *Journal of public policy, 6*(4), 371-398.

Worthy, B. (2015). The impact of open data in the UK: Complex, unpredictable, and political. *Public Administration, 93*(3), 788-805.

Wu, J. D. (2013). Incentives and outcomes: China's environmental policy. *NBER Working paper 18754*.

Xinhua Net. (2015, 11 13). *13 Cases of Pollution Data Manipulation in Shandong Province*. Retrieved from Chinese Government: http://www.gov.cn/xinwen/2015-01/13/content_2803737.htm

Xinhua Net. (2018). Retrieved from http://www.xinhuanet.com/2018-06/24/c_1123026369.htm

Xu, M. (2018, June 9). *China to launch broader environmental inspections this month*. Retrieved from Reuter: https://www.reuters.com/article/us-china-pollution/china-to-launch-broader-environmental-inspections-this-month-idUSKCN1J506F

Zheng, L., & Na, M. (2020). A pollution paradox? The political economy of environmental inspection and air pollution in China. *Energy Research & Social Science, 70*(101773).

12 Appendix

Appendix A. Spatial weight matrix

For testing the hard misreporting using DID estimator, I allow the spatial autocorrelation of the dependent variable. In addition to the continuous distance weight matrix where longer distance from each station, the smaller the weight, I also test the robustness of the results to different spatial weight matrices. I use a threshold distance matrix. For the threshold distance matrix, different distance thresholds have been experimented with, and the one that is finally being used is a threshold distance matrix of 5 km as it corresponds to the minimum distance necessary to connect 95% of the stations to at least one neighbor. In other words, all stations that have their location within 5 km are considered as neighbors. This weight matrix also results in ten stations with no neighbor, and they are excluded from the analysis. The corresponding spatial weights have the following form:

$$w_{ij} = \begin{cases} 1, & 0 \leq d_{ij} \leq 5 \text{ km} \\ 0, & d_{ij} \geq 5 \text{ km} \end{cases}$$

The following graph shows the treatment effect using a distance weight matrix. The results are largely consistent: treated stations experienced a substantial increase in the ability to translated AOD into reported pollution in the post-treatment periods.

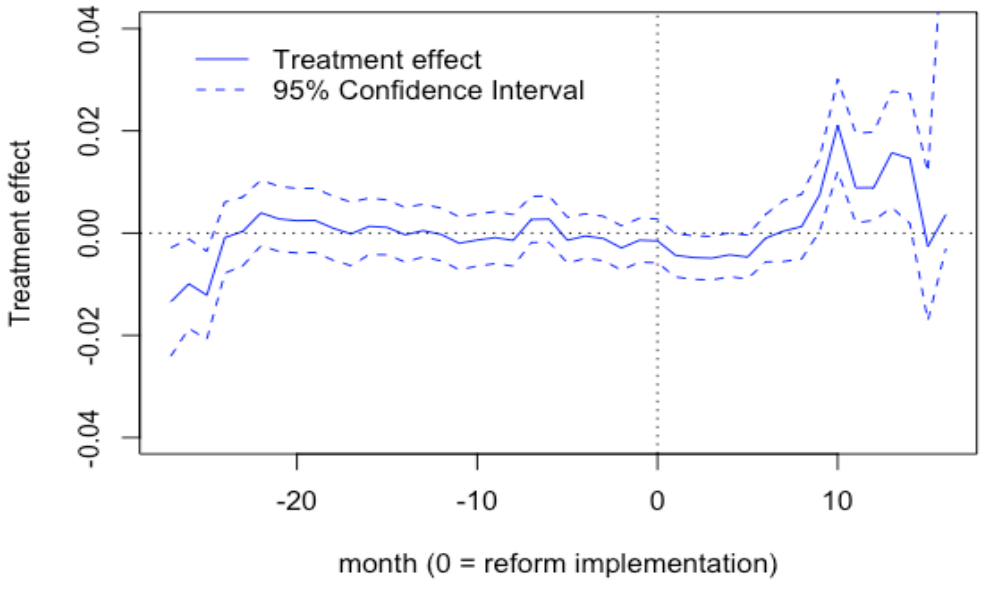


Figure 3-20 Treatment effect by month (Hard misreporting, distance weight matrix)

Appendix B. Discussion on the cause of misreporting

This paper focuses on decentralized versus centralization environmental management system and its impact on misreporting. This part of the appendix section offers a summary of the cause of government misreporting, which is related but out of the scope of this paper. Previous studies have pointed out that high political incentive leads to a distortive effect. The starting point of the analysis is that governments of all types always have the motives to exaggerate their performances or achievements (Oates & Schwab, 1988), despite the fact that incentive structure in democratic bureaucracies can be different from that in authoritarian bureaucracies like China. In democratic countries, the incentive comes from attracting voters and outperform the opponents in the election, while in authoritarian countries, the incentive comes from standing out in performance evaluation carried out by upper-level supervisors to win higher promotion chances (Wu, 2013). The canonical principal-agent model has argued the controversy in using performance incentives in bureaucracies (Besley & Prat, 2006) characterized by the tradeoff between motivating agents and distorting their efforts (Hölmstrom, 1979). If the incentive is strong while the information is asymmetric (or lacking verifiability) between principal (either voters or upper-level authorities) and agent (politicians or local governments), information manipulation is more likely to occur (Mookherjee, 2015). The study also shows that information distortion is more likely to happen in difficult observe objectives. This is especially the case for environmental data which are harder to observe (since environmental quality involves many indices) and are harder to detect manipulation (since they do not interrelate with other statistics, a counterexample is economic data, which is more difficult to manipulate because they need to be consistent with multiple sources and accounts). Sappington (1991) offers a nice review of some seminal theoretical studies.

Empirical studies have found support to these arguments. Ma & Zheng (2018) test the distortive effect of a mandate energy-saving policy in China, which has veto power on local leader promotion. test the relationship between the issuing of a mandate energy-saving policy in China, which has one-vote veto power on local leader promotion, and the extent of biased energy data. Their results suggest the distortive effect of strong political incentives. Similarly, Fisman & Wang (2017) study the perverse effect of a program designed to reduce accidental deaths. Their results show a sharp discontinuity in reported deaths at the death ceiling, suggestive of manipulation. Acemoglu et al studies misreporting behavior of Columbian colonel when facing high power political incentive. In this case, innocent civilians were killed and misrepresented as guerillas

during counterinsurgency. Their results show that there were significantly more false positives during the period of high-powered incentives in municipalities especially those with weaker judicial institutions. A similar example is the exaggeration of body counts in Vietnam (Bohte & Meier, 2000). Finally, Kalgin (2016) studies the distorted Russian statistics and attribute such misreporting to the blame avoidance attempt of local government.