The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats*

Oriana Bandiera, Michael Carlos Best, Adnan Qadir Khan, & Andrea Prat

June 2021

Abstract

We design a field experiment to study how the allocation of authority between frontline procurement officers and their monitors affects performance both directly and through the response to incentives. In collaboration with the government of Punjab, Pakistan, we shift authority from monitors to procurement officers and introduce financial incentives in a sample of 600 procurement officers in 26 districts. We find that autonomy alone reduces prices by 9% without reducing quality and that the effect is stronger when the monitor tends to delay approvals for purchases until the end of the fiscal year. In contrast, the effect of performance pay is muted, except when agents face a monitor who does not delay approvals. Time use data reveal agents' responses vary along the same margin: autonomy increases the time devoted to procurement and this leads to lower prices only when monitors cause delays. By contrast, incentives work when monitors do not cause delays. The results illustrate that organizational design and anti-corruption policies must balance agency issues at different levels of the hierarchy.

*Bandiera: London School of Economics o.bandiera@lse.ac.uk, Best: Columbia University and NBER michael.best@columbia.edu, Khan: London School of Economics a.q.khan@lse.ac.uk, Prat: Columbia University andrea.prat@columbia.edu. This experiment was preregistered in the Social Science Registry at https://www.socialscienceregistry.org/trials/610. Its pre-analysis plan and a "populated" version of the pre-analysis plan (Duflo *et al.* , 2020) are available there also. We are grateful to the World Bank, the International Growth Centre, and the JPAL Governance Initiative for financial support. We thank Ahsen Omar Majid, Sher Afghan Asad, Maha Rehman, Ameera Jamal, Omar Gondal, Khawaja Hussain Mahmood, Subhan Khalid, Sophia Tahir Mir, Ovais Siddiqui, Noor Sehur, Zain ul Abideen, Ahsan Farooqui, Natasha Ansari and Reem Hasan for outstanding research assistance in Lahore, and Hamza Husain, Sakshi Gupta and Advitha Arun for outstanding research assistance at Columbia. We thank the Government of Punjab and in particular Naeem Sheikh, Imdad Bosal, Umer Saif, Zubair Bhatti and Ali Bahadar Qazi for their collaboration over the years. We thank many seminar participants and especially Abhijit Banerjee, Ernesto Dal Bo, Bob Gibbons, Imran Rasul, and Guido Tabellini for helpful comments and discussion. All errors are our own.

1 Introduction

Organizations bring together people with different interests, information and skills to work towards a common goal. To achieve this, organizations make two interdependent choices: how to allocate decision making rights to agents at different layers of the organization's hierarchy, and how to monitor and motivate their behavior.

Organization theory, from the foundational work of Coase (1937) and Simon (1951) to the recent contributions reviewed by Bolton & Dewatripont (2013) and Gibbons & Roberts (2013), points to the allocation of authority as one of the choices at the core of organization design. By contrast, field work, guided by the single-layer principal-agent framework, tends to focus on performance rewards, while holding the architecture of the organization fixed (see, e.g. Bandiera *et al.* , 2011; Finan *et al.* , 2017, for reviews).

This paper brings the two design choices—incentive provision and authority allocation—together by means of a large-scale field experiment conducted in collaboration with the government of Punjab, Pakistan. Our context is public procurement, an activity that represents approximately 12% of GDP in the average OECD country, and which is notoriously subject to agency problems: Procurement officers are tasked with buying goods they do not use with money they do not own (Laffont & Tirole, 1994) and they operate in an environment characterized by contractual incompleteness and high transaction costs (Bajari & Tadelis, 2001). How best to tackle this is the subject of intense debate, with one camp strongly in favor of strict rules and intense monitoring (OECD, 2009) and the other arguing in favor of simplification and autonomy (Kelman, 1990). We study how the allocation of authority between officers and their monitors, who face their own agency issues, determines performance.

A simple framework illustrates how procurement outcomes depend on incentives and the allocation of authority between officers and monitors. Both agents are defined by a type that determines whether they are aligned with the organization. The equilibrium price is a function of the strength of incentives and the officers' and monitors' types. If these are equal, shifting authority from the monitor to the officer lowers prices because it eliminates the "competing bandits" problem (Shleifer & Vishny, 1993). A fortiori, prices will fall whenever the officer is better aligned and will rise only if the monitor is better aligned. This is where the complexity of the organization comes into play: In the simple principal-agent model the monitor is perfectly aligned with the principal and would not impose inefficient monitoring costs on the organization.

Performance pay for the officer always decreases prices but the effect size depends on the monitor's type. If he is misaligned, the officer cannot do much to reduce prices as these are mostly kept high by the monitor.

To create variation in the policy parameters we randomly allocate 600 procurement officers (POs) to four groups: a control group, an autonomy group, a pay for performance group and a group that gets both. The autonomy treatment shifts decision making rights from the monitors to the POs by removing the monitor's discretion over the list of documents that they can demand as part of the audit, and by giving POs full decision rights over purchases in cash up to 10% of the average PO budget. The pay for performance treatment is a rank order tournament within district and administrative department which pays prizes ranging from half a month's salary to two months' salary on the basis of value for money.

Our main measure of procurement performance is price conditional on quantity and the precise nature of the good being purchased, including delivery speed and transport costs. To maintain comparability we focus on a sample of over 20,000 purchases of generic goods gathered through an online reporting system we developed to collect detailed information on the attributes of each purchase.[1] The experiment spanned two years and 26 districts, allowing us to construct a proxy for each monitor's type since each district has its own monitors.

Our findings are as follows. First, consistent with the fact that procurement officers are given orders to fill based on the needs of the organization, the treatments do not affect the composition, quantity or attributes of the items purchased. Second, autonomy reduces prices by 9% on average either on its own or in combination with performance pay. Performance pay on its own reduces prices by 3% but we cannot reject the null that the effect is equal to zero. Our findings are consistent with, and provide micro foundations for, the result that autonomy, but not incentives, is correlated with performance in bureaucracies (Rasul & Rogger, 2018; Rasul *et al.* , 2019), that autonomous schools have better performance (Bloom *et al.* , 2015a,b) and that reducing discretion in environmental inspections increases costs without reducing pollution (Duflo *et al.* , 2018).

To benchmark the effects we compare the savings from our treatments to the cost of public goods. Our point estimates suggest that the savings from the autonomy treatment from the relatively small group of offices in our experiment are sufficient to fund the operation of five schools or to add 75 hospital beds. This is twice the savings from the combined treatment and six times the savings from the incentive treatment. Despite

[1] Despite the fact that each purchase of these generic goods is small, cumulatively they account for a large share of procurement expenditures. As table 1 shows, they account for 53% of a typical office's budget.

the modest savings, the rate of return on the incentives treatment is 45% since the small per-purchase savings are applied to a large base of expenditure.

These overall effects can be reconciled by viewing the results through the lens of our model. When monitors are relatively misaligned our model predicts that autonomy will improve performance but incentives will not. Moreover, the model provides predictions for how the treatment effects will vary with the types of the agents. We allow the effects to vary with the monitor's type, which we measure with the share of transactions approved at the very end of the fiscal year (Liebman & Mahoney, 2017). This captures both inefficiency, i.e. a slow monitor, and corruption, i.e. a monitor who holds officers up until their budget lapses. We find that performance pay reduces prices by 6% when the monitor approves transactions quickly over the year while the effect goes to zero when the monitor holds up more than 48% of transactions until the end of the fiscal year. The effect of autonomy has the opposite pattern: it is zero when the monitor is "good" and it reduces prices up to 20% when the monitor is "bad". While this analysis was not pre-specified and hence should be treated as more exploratory, our model provides a simple and compelling framework within which to rationalize the divergent findings on the overall effects, and one whose auxiliary predictions are borne out in the data.[2]

Time use data reveal that the officers' response to treatment follows a similar pattern: all treated officers devote more time to procurement, but those in the incentive group only do so when the monitor is "good", while those in the autonomy group put in extra time when the monitor is "bad". The experimental design also allows us to measure the impact of officers on procurement outcomes: A back of the envelope calculation indicates that these changes in time use explain 72% of the effect of incentives and 62% of the effect of autonomy. In addition, we find that shifting autonomy from "bad" monitors to officers reduces delays and the likelihood that the monitor waits until the very end of the year to approve a purchase.

Taken together the results indicate that the two policy instruments are effective under different circumstances: giving autonomy to the agent is desirable when it means taking it away from an extractive monitor, while incentives are ineffective in this case because the agent has limited control over prices, and vice versa. In line with this, the effect of the combined treatment always falls between the other two.

Our findings point to the importance of understanding the drivers of bureaucrats' behavior when seeking to improve performance in the public sector. The findings echo the cross-country patterns documented in Bosio *et al.* (2020), who show that laws that con-

[2]This experiment was preregistered in the Social Science Registry at `https://www.socialscienceregistry.org/trials/610`. Its pre-analysis plan and a "populated" version of the pre-analysis plan (Duflo *et al.* , 2020) are available there also.

strain procurement outcomes are effective in most low income countries but harmful in richer countries. Policies based on the assumption that most bureaucrats are corrupt are likely to backfire when this is not in fact the case, for instance by distorting incentives to undertake socially optimal actions for fear of reputational damage (Leaver, 2009), or of being punished for breaking the rules (Shi, 2008). The results also speak to recent studies that use observational variation to show how anti-corruption measures such as audits are ineffective or even detrimental once the response of private sector agents is taken into account (Yang, 2008; Gerardino *et al.* , 2017; Lichand & Fernandes, 2019). Our paper also contributes to the experimental literature measuring the organization of corruption (Olken & Barron, 2009; Sanchez de la Sierra & Titeca, 2019; Weaver, 2020). Finally, our paper contributes to the debate on the optimal amount of discretion in procurement (Szucs, 2017; Coviello *et al.* , 2018).

The remainder of the paper proceeds as follows. In section 2 we present the empirical context for our experiment, and section 3 describes the experimental design. Section 4 develops the conceptual framework we use to guide our empirical analysis. Section 5 presents our results, and our conclusions are in section 7.

2 Context and Data

2.1 Procurement in Punjab

Our study takes place in the province of Punjab, Pakistan, home to 110 million people, and divided into 36 administrative districts. Our study took place in 26, covering 80% of the population and the largest districts.[3] Each government office has one employee who is designated as the Procurement Officer (PO). He or she wields the legal authority to conduct small and medium sized procurement purchases.[4] Offices are allocated budgets under a range of accounting heads (salary, repairs, utilities, etc.)—including procurement—and are not permitted to move budget across categories with very limited exceptions. Before making payments to vendors, the POs are required to submit their purchases for pre-audit approval by an independent agency of the federal government known as the Accountant General's office (AG). The AG has offices in each of the districts of the province, monitoring the purchase of offices in that district.

A typical procurement process for the purchase of a generic item like the ones we

[3]These districts were chosen on the basis of logistical feasibility, being geographically contiguous and ruling out the most remote districts. Appendix figure A.5 shows the location of the offices.

[4]The title of this position is known as the "Drawing and Disbursement Officer" of the office.

study proceeds in five steps, as summarized in panel A of figure 1. First, an employee of the office makes a request for the purchase of an item (for example, a teacher might request the purchase of pens for the classroom). Second, the PO approves the purchase and surveys the market for vendors who can supply the required item and solicits quotes for the item. Once the PO has received enough quotes for the item, he/she chooses which vendor to allocate the contract to.[5] Third, the vendor delivers the items to the public body and the PO verifies receipt of the items. Fourth, the PO prepares the necessary documentation of the purchase and presents it to the AG office.

Fifth, the AG reviews the paperwork. The remit of the AG is to check that everything has been done in accordance with the rules. If the AG is satisfied with the documentation, they sanction the payment and gives the PO a check made out to the vendor. If the AG is not satisfied, they can demand more thorough documentation that the purchase was made according to the rules. This ability to withhold approvals and the ensuing delay are the key source of the AG's power over POs. Ultimately, POs who are found to be in breach of the rules are punished.[6] Indeed in our survey POs say that not following rules is the main threat to their career.

2.2 Measuring Bureaucratic Performance

The government of Punjab considers that the primary purpose of public procurement is to ensure that "...the object of procurement brings value for money to the procuring agency..." (Punjab Procurement Regulatory Authority, 2014). In line with this, we developed a measure of bureaucratic performance that seeks to measure value for money in the form of the unit prices paid for the items being purchased by POs, adjusted for the precise variety of the item being purchased. The backbone of our approach is collecting detailed data on the attributes of the items being purchased with which to measure the precise variety of the items.

We proceed in two steps. First, we restrict attention to homogeneous goods for which we could gather detailed enough data to adequately measure the variety of the item being

[5]For very small purchases, only one quote is needed. For most of the purchases we consider, POs must obtain three quotes and then choose the cheapest one.

[6]The institutional structure involving the AG and audit mechanisms that we study in Punjab are similar to ones in many other countries where independent, constitutionally appointed bodies like the AG audit government spending imposing all kinds of disciplinary measures ranging from resolution of minor discrepancies to major punishments like dismissal from service, lodging of criminal cases and recovery of losses to government. This, and similar systems, are ultimately overseen by the Public Accounts Committee of the parliament. In theory, the audit process has all these powers and just the threat of being under inquiry initiated by the AG can be damaging for officers as it hurts their career concerns - they cannot be promoted while facing an inquiry.

purchased (similar to the approach taken in Bandiera *et al.* 2009 and Best *et al.* 2019).[7] Second, we partnered with the Punjab IT Board (PITB) to build an e-governance platform—the Punjab Online Procurement System (POPS). This web-based platform allows offices to enter detailed data on the attributes of the items they are purchasing. We trained over a thousand civil servants in the use of POPS and the departments we worked with required the offices in our experimental sample to enter details of their purchases of generic goods into the POPS system. To ensure the accuracy of the data we randomly visited offices to physically verify the attributes entered into POPS and collect any missing attributes required.[8]

After running the POPS platform for the two years of the project and cleaning the data the officers entered, our analysis dataset consists of the 25 most frequently purchased goods—a total of 21,503 purchases. Dropping the top and bottom 1% of unit prices results in a dataset of 21,183 observations.[9] Figure 2 shows summary statistics of the purchases in the POPS dataset. The 25 items are remarkably homogeneous goods such as printing paper and other stationery items, cleaning products, and other office products. While each individual purchase is small, these homogeneous items form a significant part of the procurement budgets of our offices. As table 1 shows, generic goods are 53% of the typical office's budget.

Despite the homogeneous nature of the items being purchased, prices are quite different. Figure 2 shows this variation for each product, and figure A.1 shows the joint distribution of prices paid and the standardized price of each purchase (a measure of the item's variety described in section 5.1 that can be interpreted as the predicted expected price if the item had been purchased in the control group). Both figures display variation in prices, even for items of the same variety, suggesting different bureaucrats are paying different amounts for identical products. This degree of price dispersion for very homogeneous goods is not uncommon in the public sector, similar levels have been documented in the United Kingdom (National Audit Office, 2006), Italy (Bandiera *et al.* , 2009) and Russia (Best *et al.* , 2019).

To elicit procurement officers' perceptions of their incentives to perform procurement well, we asked officers what types of errors would be detrimental to their career progress. Since civil servants in Punjab are not typically paid based on their performance, the main

[7]To do this, we chose accounting codes from the government's chart of accounts that we expected to contain mostly or exclusively generic goods. The list of accounting codes is contained in appendix table A.1.

[8]Somewhat surprisingly, our random audits did not uncover any instances of misreporting of goods' attributes.

[9]The majority of these outliers are the result of officers adding or omitting zeros in the number of units purchased.

incentive they face is that their performance is considered when decisions are made on their postings and on their progression up the civil service hierarchy. Specifically, two of the options we asked officers about are how detrimental overpaying in their procurement purchases would be, and how detrimental failing to complete the required documentation would be. Appendix figure A.2 shows the results. While the officers respond that both transgressions would be detrimental for their careers, they report that having incomplete documentation is a severe impediment much more often than overpaying. This stands in clear contrast to the government's stated goal when conducting public procurement—to achieve value for money (Punjab Procurement Regulatory Authority, 2014), and motivates our two treatments.

3 Experimental Design

3.1 Design of Experimental Treatments[10]

In the status quo, the authority to approve purchases and pay vendors lies with the Accountant General (AG). Our *autonomy* treatment shifted decision-making power over which documents can be required in order to issue a payment to a vendor away from the AG. To achieve this, we conducted focus groups with Procurement Officers (POs) and their staff to elicit their demand for policy changes that could empower them to achieve greater value for money. We then brought their proposals to the government and reached an agreement on which policy changes to implement.[11]

Our treatment altered the procurement process to limit the AG's power in two ways. First, we offered each PO a cash balance of Rs. 100,000 (USD 1,000), over which they had full authority. That is, they could use this money to make payments to vendors without having to seek pre-audit approval from the AG, thus completely removing the AG's authority over the documentation of this part of the office's spending, as illustrated in the top path in panel B of figure 1.[12]

[10]This experiment was preregistered in the Social Science Registry at `https://www.socialscienceregistry.org/trials/610`. Its pre-analysis plan and a "populated" version of the pre-analysis plan (Duflo *et al.* , 2020) are available there also.

[11]The importance of these policy changes is confirmed in our endline survey. Figure A.3 shows the responses the control group gave when asked to allocate 100 points between a set of potential reasons for the lack of value for money in public procurement. The three most important reasons are that budgets are released late, that POs do not have enough petty cash to make purchases quickly, and that the AG's requirements are not clear.

[12]Petty cash is still subject to all the same legal scrutiny and documentary requirements as ordinary spending during post audit after the conclusion of the financial year. The only difference is that it does not require pre-audit approval by the AG.

Second, we created and distributed a checklist of the documents that the AG can lawfully require in order to approve a purchase, even when the payment is not to be made with petty cash, as shown in the bottom path in panel B of figure 1. The list limits the AG's authority to decide which documents are required for payment by restricting them to the documents in the checklist. The finance department endorsed and sent the checklist to the offices, making it a credible signal of what the requirements were. The AG was also informed by the finance department that these were the requirements it wanted the AG to check during pre audits.[13]

Giving more autonomy to procurement officers can improve outcomes by reducing payment delays, allowing them to buy from a wider range of vendors and generally avoid mark-ups imposed by the AG. Autonomy, however, also makes it easier for POs to embezzle funds and limits the AG's discretion in identifying and combatting new loopholes POs may attempt to exploit to circumvent procurement rules. Finally, while our treatment is tailored to the institutional context, it is easily adaptable to any situation in which an agent's decision making power is constrained by another agent.

Our *incentives* treatment aligned POs' incentives with the government's by providing them with financial incentives to improve value for money. Officers' performance was evaluated by a committee established for this purpose. The committee was co-chaired by the President of the Institute of Chartered Accountants Pakistan (ICAP), a well-respected, senior, private-sector monitor, and the director of the Punjab Procurement Regulatory Authority (PPRA). Delegates from each of the line departments, the finance department, and the research team rounded out the committee. Based on common practice in the private sector, the committee was tasked with ranking the procurement officers' performance by applying a wholistic assessment to the officer's performance at achieving the aims of public procurement. To seed the discussions, the research team provided an initial ranking of the procurement officers according to our measure of value added described in section 2.2, though the committee were told they had absolute freedom to alter the ranking.

Based on the committee's ranking, bonuses were paid. The *gold* group, comprising the top 7.5% of officers, received two months' salary. The *silver* group, the next 22.5% of officers, received one month's salary. The *bronze* group, the next 45% of officers, received half of a month's salary. Finally, the remaining 25% of officers did not receive an honorarium. The committee met twice a year. Based on the interim rankings at the middle of the year, officers received payments of half of the bonus amounts, which were then credited against the bonuses received in the final ranking at the end of the year.

[13]To increase the power of these treatments, a third component attempted to improve the frequency and regularity of budget releases. However, as we document in appendix figure A.4, it was not possible to implement this.

We made several design choices to increase the salience, credibility and feasibility of this treatment that are worth noting. First, we chose a form of incentive that is allowed under the existing rules so that it is both feasible and easily scaleable should the government choose to do so. Second, we chose a prize structure that meant that 75% of officers received a prize. Third, we chose to have the committee meet twice a year. Together, these meant that many POs would experience receiving a prize, and that the bonuses were salient during the second half of the year when the bulk of procurement expenditure takes place. Moreover, the incentive treatment was in place during the pilot year to build credibility so officers already had experience with the treatment when the second, focal year began.

3.2 Experimental Population and Randomization

The experiment was conducted in collaboration with several agencies of the government of Punjab. The Punjab Procurement Regulatory Authority (PPRA), the Punjab Information Technology Board (PITB), the Accountant General's (AG) office, and the finance department worked with us to design and oversee the treatments. We sampled offices from the four largest front-line departments—Higher Education, Health, Agriculture, and Communication & Works. Within these departments we sampled from offices with procurement budgets in the 2012–13 fiscal year of at least Rs. 250,000 (USD 2,500).

In June 2014, we randomized 688 offices into the four treatment arms, stratifying by district $\times$ department to ensure balance on geographical determinants of prices and the composition of demand. Offices were told by their departments that they were part of a study to evaluate the impact of policy reforms under consideration for rollout across the province and that their participation was mandatory, including entering data into the POPS system and cooperating with occasional survey team visits. With this backing, 587 offices, or 85% of the sample, participated in trainings on the POPS system and on the implications of their treatment status for how they conduct procurement.

Table 1 presents summary statistics on a range of variables in the participating offices. The table shows that the participation rate is balanced across the treatment arms, as are the vast majority of office characteristics and budgetary variables available in the finance department's administrative data. We regress each variable on dummies for the three treatments and report the coefficients along with their robust standard errors in parentheses and p-values from a randomization inference test of the null of a zero effect. For each variable we also report the F-statistic on the test that all treatments have no effect with its corresponding p-values using the asymptotic variance, and the randomization inference p-value. Of the 24 variables presented, the hypothesis that all treatments have no effect is rejected for only one variable—the number of accounting entities the office controls, and so we control for this in our estimation of treatment effects.[14]

Participating offices' compliance with the requirement to enter data into the POPS system was also balanced. Figure A.8 estimates office-level measures of POPS compliance and shows that their full distributions are balanced across treatments, while table A.3 shows that the mean compliance rate varies across accounting categories, but even controlling for this, is balanced across treatments. Overall, we conclude that the randomization produced a balanced sample and that compliance was high and balanced across the treatment arms.

Table A.2 summarizes the timeline of the project. The 2014–15 fiscal year was the pilot year for the project. The POs were informed of the project and introduced to POPS. All POs were invited to receive training on the use of POPS and to start entering data into the system. The incentives treatment was in place so that the members of that treatment group would experience receiving the bonuses, but the autonomy treatment was not.[15] Then, in year 2 (the 2015–16 fiscal year), the autonomy treatment was also rolled out. The experiment ended at the end of June 2016, following which we conducted an endline survey and gathered missing data.

4 Conceptual Framework

The literature on the "organization of corruption" (Shleifer & Vishny, 1993; Guriev, 2004; Banerjee *et al.* , 2012) studies situations where multiple potentially corrupt agents may be involved in a public deliberation. The ultimate outcome is determined by the motivations of the individual agents as well as the organizational architecture of the deliberation process. The literature studies equilibrium behavior under different architectures and suggests ways to design institutions that are more robust to the risk of corruption.

Inspired by this literature, we consider a highly stylized procurement process where a public body must buy one unit of a good of a given quality. Two agents may be involved in the process: a purchasing officer (the PO) and a monitor (the accountant general, AG). We consider two arrangements: an autonomous PO making decisions on his own (autonomy); and a PO making decisions that are subject to the veto power of an AG

[14]This is likely to have occurred because the office that controls a small number of accounting entities was incorrect in the administrative data used for the randomization. When this occurred, we assigned accounting entities to the treatment received by their actual office. Since offices with more accounting entities have a greater chance of having one incorrectly recorded, this can lead to this imbalance.

[15]Discussions between the research team and the government about the precise nature of the treatment and how to implement it were still ongoing.

(non-autonomy).

Both the PO and the AG can be "aligned" or "misaligned". Aligned agents behave in the interest of the citizens. This might be due to some form of intrinsic motivation, or from differential career concerns. Misaligned agents instead behave sub-optimally, either because they are corrupt and they prefer higher prices in exchange for bribes or other favors (active waste), or because they are lazy and their desire to minimize effort results in inefficient processes that lead to higher prices (passive waste).

Our model can be interpreted within the framework developed by Bosio *et al.* (2020), where a PO faces multiple suppliers, one of whom is an insider. Even though the good is sold with a competitive mechanism, the PO has some latitude in excluding a supplier. A non-aligned PO may accept a bribe to exclude all suppliers but the insider, thus driving up the auction price. We simplify Bosio *et al.* (2020) by holding quantity constant and we extend by adding an agent, the AG. The exclusion decision would be made by the PO in the autonomy case or be a function of the individual actions of the PO and the AG in the non-autonomy case. An example of passive waste is a public body that acts so slowly that suppliers demand higher prices because they know they will be paid late. While the present model makes assumptions directly on equilibrium prices, Appendix C offers a micro-founded version that starts from the utility functions of the PO and the AG.[16]

Under the autonomy arrangement, the price paid depends on whether the PO is aligned (the price is p_A) or misaligned (p_M), with the assumption that $p_M > p_A$. In the non-autonomy case, the outcome depends on the type of both agents: p_{AA} (both PO and AG are aligned), p_{AM} (aligned PO, misaligned AG), p_{MA} (misaligned PO, aligned AG), and p_{MM} (both misaligned). Agent type has the expected monotonic effect. Better types reduce the purchase price: $p_{AA} < p_{AM} < p_{MM}$, and $p_{AA} < p_{MA} < p_{MM}$. Under both arrangements, good POs obtain the best possible price: $p_{AA} = p_A = c$, where c is the minimum possible price at which the supplier is willing to sell.[17]

We also make four assumptions about the agency relationship between the AG and PO:

1. *Good Monitor Effect*: $p_{MA} < p_M$. A good monitor has a positive discipline effect on a bad agent: an aligned AG makes a misaligned PO behave better than he would if he was not monitored (either because she makes him less corrupt or more efficient).

2. *Bad Monitor Effect*: $p_{AM} > c$. The bad monitor has a detrimental effect on good

[16] The micro-foundation in Appendix C also goes beyond the binary types used here. Both the PO and the AG have continuous types and therefore the model will give rise to a continuum of equilibrium prices.

[17] One could also assume that imposing an aligned AG on top of an aligned PO would add some red tape anyway, therefore leading to $p_{AA} > p_A$. The key results are unchanged if such effect is smaller than the effect of imposing a misaligned AG on top of an aligned PO – a natural assumption.

agents: a misaligned AG increases the price obtained by a good PO. In the active waste case, the negative effect may come from demanding a bribe for authorizing the purchase. In the passive waste case, it comes from slowing down the process, thus leading the supplier to demand a higher price.

3. *Competing Bandits Effect*: $p_{MM} \geq p_M$: This is the corruption effect identified by Shleifer & Vishny (1993). The amount of corruption is increasing in the number of corrupt agents with veto power, the rationale being similar to the industrial organization effect that double marginalization has on price in the sale of complementary products. In the present setting, adding a misaligned AG on top of a misaligned PO will lead to higher prices.

4. *Complementarity Between Types*: $p_{MM} - p_{AM} \leq p_{MA} - c$. There is some complementarity between the types of the two agents. The price reduction due to having an aligned PO rather than a misaligned PO is smaller if the AG is misaligned. This is because a misaligned AG is more likely to appropriate any benefit created by an aligned PO. For instance, if the PO decreases his bribe demand, a misaligned AG will increase hers.[18]

The two experimental treatments can be interpreted within the model. The incentive treatment provides POs with better alignment: we assume that all POs behave like aligned POs. The effect of the autonomy treatment is to take the AG out of the picture, thus moving from the non-autonomy arrangement to the autonomy arrangement. The combined treatment performs both operations at the same time.

Let the share of POs that are misaligned be θ_{PO} and the share of AGs that are misaligned be θ_{AG}. We obtain the following average prices

	No Incentive	Incentive
No Autonomy	$\theta_{PO}\theta_{AG}p_{MM} + \theta_{PO}(1-\theta_{AG})p_{MA} + (1-\theta_{PO})\theta_{AG}p_{AM} + (1-\theta_{PO})(1-\theta_{AG})c$	$\theta_{AG}p_{AM} + (1-\theta_{AG})c$
Autonomy	$\theta_{PO}p_M + (1-\theta_{PO})c$	c

The effect of the autonomy treatment on the average price is

$$\Delta_A = \theta_{PO}\theta_{AG}(p_M - p_{MM}) + (1-\theta_{PO})\theta_{AG}(c - p_{AM}) + \theta_{PO}(1-\theta_{AG})(p_M - p_{MA})$$

[18] For our comparative statics results, we require the good- and bad-monitor effects to hold as strict inequalities. Instead the conditions for the competing bandit effect and the complementarity between agent types can also be equalities.

and it can be decomposed into three parts. The first term captures the bandit competition effect which arises in a share $\theta_{PO}\theta_{AG}$ of monitor-officer pairs. Autonomy eliminates this and reduces prices ($p_M - p_{MM} < 0$). The second term captures the effect of having a bad monitor control a good officer. This occurs in a share $(1-\theta_{PO})\,\theta_{AG}$ of pairs and eliminating it reduces prices as $c - p_{AM} < 0$. The third term captures the effect of having a good monitor control a bad officer. This occurs in a share $(1-\theta_{AG})\,\theta_{PO}$ of pairs and eliminating it increases prices as $p_M - p_{MA} > 0$.

The overall effect can be positive or negative depending on the relative ratio of aligned AGs and aligned POs:

Proposition 1. *The autonomy treatment increases the expected price paid if and only if the AG is relatively more aligned than the PO. The condition is*

$$\theta_{AG} < \bar{\theta}_{AG}\left(\theta_{PO}\right),$$

where $\bar{\theta}_{AG}\left(\theta_{PO}\right) > 0$ is the solution to:

$$\frac{\theta_{AG}}{1-\theta_{AG}} = \frac{\theta_{PO}\left(p_M - p_{MA}\right)}{-\left(\theta_{PO}\left(p_M - p_{MM}\right) + \left(1-\theta_{PO}\right)\left(c - p_{AM}\right)\right)}.$$

Proposition 1 captures a basic intuition: having a monitoring system is a good idea if and only if the monitor is on average "better" than the monitored. An aligned monitor keeps prices down through the discipline mechanism that she imposes on misaligned POs (the good monitor effect). A misaligned monitor inflates prices through the introduction of a bribe if the PO was aligned and did not demand one (the bad monitor effect) and double marginalization if the PO was misaligned and was already asking for a bribe (the competing bandit effect). The net effect is found by comparing the probability that the AG is aligned with the probability that the PO is aligned.[19]

Note that if the AG is as aligned as the PO, autonomy is always a good idea:

Corollary 1. *If the AG is as aligned as the PO ($\theta_{AG} = \theta_{PO}$), the autonomy treatment strictly decreases the expected price paid.*

The corollary is due to the competing bandits effect. Imposing a monitor of the same quality as the monitored is not going to help and can only hurt through double marginalization. In order for monitoring to reduce prices it must be that the AG is discretely more aligned than the PO.

[19] If the bandit competition effect is absent and the the bad monitor effect is relatively low, then $\bar{\theta}_{AG}\left(\theta_{PO}\right) < 1$.

Turning to the second treatment, incentivizing the PO can never hurt as it will always motivate him to reduce the price. However, the effect is dampened if the PO is monitored by a misaligned AG.

Proposition 2. *The incentive treatment always decreases the average price paid but the size of the effect is smaller if the AG is more likely to be misaligned ($\theta_{AG} \to 1$) and goes to zero if a misaligned AG has full control over the price ($p_{MM} = p_{AM}$).*

The three theoretical results can be visualized in Figure 3. The plot depicts the effect of the two treatments as a function of the types of the two agents. Proposition 1 means there are parameter values where autonomy is beneficial and where autonomy is detrimental – and an area in between where its effect is negligible. The Corollary means that along the 45° line where the two agents have the same type, autonomy cannot be detrimental: if both agents are honest, the treatment does not do much; if both agents are sufficiently dishonest, the treatment is beneficial. Proposition 2 implies that incentives can be beneficial – if the AG is sufficiently honest – or have a negligible effect.

If the two treatments are combined, the effect is determined by the individual effects, but there may be synergies. One can show that the price reduction of the combined treatment is at least as large as the price reduction of each individual treatment (See Proposition 3 in Appendix B for a formal statement). The combination of the two treatments must be as good as the better of the two individual treatments because the incentive treatment never hurts when the PO is on his own and, if it helps, it helps at least as much as it would when the AG is present given that the AG is not treated. There may be some strict complementarity if both agents are misaligned, as the incentive treatment becomes more effective once any price reduction generated by the PO cannot be undone by the AG.

5 Procurement Performance

This section analyzes the overall impacts of the experiment on bureaucratic performance. The main task of a procurement officer is to receive requests for goods from his/her colleagues and purchase them at a good price. Therefore, *a priori* we do not expect other aspects of procurement performance to be affected by the treatments since the demand for the good is coming from a different officer than the person in charge of procurement. Nevertheless, we investigate the impact of the treatments on a range of procurement performance outcomes.

5.1 Measuring Good Varieties

To be able to isolate the effects of the treatments on the prices procurement officers pay, we need to be able to compare purchases of exactly the same item. Otherwise, we risk conflating differences in the precise variety of the goods being purchased with the prices paid for them. Moreover, the treatments may have affected the varieties of goods POs purchase and these are treatment effects we are interested in in their own right.

The goods in our sample are chosen precisely because they are extremely homogeneous. Nevertheless, there may still be some differentiation across items and so we use four measures of the variety of the goods being purchased. First, we use the full set of attributes collected in POPS for each good. This measure has the advantage of being very detailed, but comes at the cost of being high-dimensional. Our three other measures reduce the dimensionality of the variety of controls. To construct our second and third measures, we run hedonic regressions using data from the control group to attach prices to each of the goods' attributes. We run regressions of the form

$$p_{igto} = \mathbf{X}_{igto}\lambda_g + \rho_g q_{igto} + \gamma_g + \varepsilon_{igto} \tag{1}$$

where p_{igto} is the log unit price paid in purchase i of good g at time t by office o, q_{igto} is the quantity purchased, γ_g are good fixed effects, and $\mathbf{X}_{igto}$ are the attributes of good g.

Our second, *"scalar"* measure of good variety uses the estimated prices for the attributes $\hat{\lambda}_g$ to construct a scalar measure $v_{igto} = \sum_{j\in A(g)}\hat{\lambda}_j X_j$ where $A(g)$ is the set of attributes of item g. v_{igto} can therefore be interpreted as the expected price paid for a good with these attributes if purchased by the control group. Our third, *"coarse"* measure studies the estimated $\hat{\lambda}_g$s for each item and partitions purchases into high and low price varieties based on the $\hat{\lambda}_g$s that are strong predictors of prices in the control group. Finally, our *"machine learning"* measure develops a variant of a random forest algorithm to allow for non-linearities and interactions between attributes that the hedonic regression (1) rules out. Appendix D provides further details.

5.2 Identification

To estimate the treatment effects on bureaucratic performance we estimate equations of the form

$$y_{igto} = \alpha + \sum_{k=1}^{3}\eta_k \text{Treatment}_o^k + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto} \tag{2}$$

where y_{igto} is the outcome of interest in purchase i of good g at time t by office o; q_{igto} is the log quantity purchased, capturing good-specific bulk discounts; δ_s and γ_g are stratum and good fixed effects, respectively; and $\mathbf{X}_{igto}$ are purchase-specific controls. We weight regressions by expenditure shares in the control group so that treatment effects can be interpreted as effects on expenditure, and the residual term ε_{igto} is clustered at the cost centre level.[20]

The random allocation of offices to treatments means that the coefficients η_k estimate the causal effect of treatment k on unit prices under the assumption of stable unit treatment values (SUTVA) (Rubin, 1980; Imbens & Rubin, 2015). This might be violated if, for example, the AG extracts more from the offices in the control group because it is more difficult to extract from offices in the autonomy treatment. In practice, this is unlikely to affect our estimates because, as shown in Appendix figure A.6, AG officers have typically fewer than 20% of their cost centres in any treatment group. We can also test SUTVA more directly by seeing whether price increases between year 1 (before the roll out of the autonomy treatment) and year 2 (after the roll out) are larger under AG offices with a larger share of offices receiving the autonomy treatment. Appendix figure A.7 shows that, if anything, the point estimate is negative, supporting the SUTVA assumption.

When we are interested in studying effects on prices we need to ensure that we are comparing purchases of exactly the same varieties of items. If, however, the treatments directly affect the varieties of items being purchased, the η_k coefficients in equation (2) with price as the outcome estimate a combination of the treatment effects on prices and on the composition of purchases.[21] With this in mind, below we directly estimate treatment effects on the varieties of items being purchased. These effects are interesting in their own right and also allow us to gauge the magnitude of the potential composition effect described above. To do this, we estimate equation (2) with our scalar, coarse, and machine learning variety measures as outcomes.[22]

Two additional concerns relating to the varieties of items being purchased may affect

[20]Cost centres are accounting entities to which budget is formally assigned. In most cases each office is a cost centre, but in some cases an office is in charge of two or three cost centres. When this happens all cost centres under the same office are allocated to the same treatment.

[21]To see this, consider a simplified version of our setting. Suppose that purchases are associated with potential prices $p(D,V)$ depending on a binary treatment $D \in \{0,1\}$ and binary good variety $V \in \{0,1\}$, and with potential quality levels $V(D)$ depending on treatment. The random assignment in the experiment implies that the potential outcomes are independent of treatment status conditional on the randomization strata S_i: $\{p_i(D,V), V_i(D)\} \perp D_i|S_i$. We can now see that a comparison of expected prices between treated and control units conditional on item type combines a treatment effect on price with a potential composition effect coming from changes in the set of purchases of high or low type in treatment versus control units $\mathbb{E}[p|D=1,V=1] - \mathbb{E}[p|D=0,V=1] = (\mathbb{E}[p(1,1)|V(1)=1] - \mathbb{E}[p(0,1)|V(1)=1]) + (\mathbb{E}[p(0,1)|V(1)=1] - \mathbb{E}(p(0,1)|V(0)=1))$ where the first term on the right-hand side is the treatment effect on price we seek to estimate, and the second is a composition effect which need not equal zero.

[22]A similar concern applies to the quantity purchased in each order. In appendix table E.1 we show that the experiment did not affect the size of each order.

our interpretation of treatment effects on prices as effects on the performance of the PO. First, POs may pay low prices but buy inappropriate goods that are ill-suited to the needs of the office they are serving. However, as we will see, there are no effects of the treatments on the varieties of items being purchased. Therefore, while the goods purchased may well be badly matched to the needs of the end users in the offices, the degree of mismatch is not affected by the treatments.

Second, changes in PO behavior may cause supply-side responses by government suppliers, in which case changes in equilibrium prices reflect both the effects of changes in demand by POs and changes in supply by vendors. While this is likely in markets in which the government is a large buyer (see, for example, Duggan & Scott Morton, 2006 for evidence that procurement affects pharmaceutical producers' private-sector prices), the products in our sample are extremely homogeneous and consumed throughout the economy, so the government's market share is likely to be small. Moreover, our experimental subjects are only part of the total demand for these products from the government.

5.3 Average Treatment Effects

We begin by studying the impact of the experiment on the prices and the varieties of goods purchased. Table 2 shows the average treatment effects estimated using equation (2) using data from the second year of the project, in which all treatments were in place. Below each coefficient we report its standard error clustered by cost centre in parentheses and the p-value from randomization inference under the null hypothesis of no treatment effect for any office in square brackets.[23] Columns 1–3 estimate treatment effects on the scalar, coarse and machine learning measures of good variety, respectively. Columns 4–8 estimate treatment effects on log unit prices paid without controls for the variety of the item purchased (column 4), controling for the full set of good attributes (column 5), and controling for the scalar (column 6), coarse (column 7) and machine learning (column 8) good variety measures.

Somewhat surprisingly, table 2 shows no evidence that the experiment affected the varieties of goods being purchased. Eight of the nine coefficients in columns 1–3 have p-values above 0.25, and in all three columns the p-value on the hypothesis that none of the treatments affected good variety in any office is insignificant at 5%. This is likely because offices' demand is relatively inelastic from year to year and because the procurement officer is charged with acquiring a particular good at a good price and has limited discretion over which variety of good is purchased.

[23] We thank Alwyn Young (2019) for producing the `randcmd` package for Stata that greatly facilitates this.

Since good varieties are not affected by the treatment, they also do not cause any bias from composition changes when estimating price effects (as discussed in section 5.2).[24] Therefore, when studying prices, we include controls for the variety of goods being purchased to improve power, but also show that our price results are robust to omitting controls for the good variety being purchased.

Turning to the treatment effects on prices, three key findings emerge from table (2). First, the point estimates of the impacts of the treatments are negative for all three treatments. However, the average impact of the incentives treatment is statistically indistinguishable from zero. This surprising finding for the incentives treatment already hints at how important it is that people who are incentivized have the autonomy to respond to the incentives they are provided, a theme we return to in section 6.

Second, the autonomy treatment reduces average unit prices paid by 8–9%, indicating that giving bureaucrats greater autonomy leads them to use it in the interests of taxpayers by procuring the goods they purchase at lower prices. Viewed through the lens of the model in section 4, this implies that the accountant general is sufficiently misaligned with the principal relative to the misalignment of the procurement officer $\left(\theta_{AG} > \bar{\theta}_{AG}\left(\theta_{PO}\right)\right)$ that removing the waste caused by complying with the monitoring activities of the accountant general more than offsets the loss of the benefits the accountant general's monitoring provides.

Third, the findings on the impact of the treatments on quality-adjusted prices paid are robust to alternative measures of the variety of good being purchased or not controlling for the goods' varieties. Intuitively, the asymptotic standard errors of the estimates are smaller when using the lower-dimensional measures of good variety as the model has more degrees of freedom. However, the p-values from randomization inference are smallest when using the full vector of good attributes as controls, consistent with the finding in Young (2019) that the benefits of using randomization inference are largest when the estimated models are high-dimensional.

Online Appendix E explores effects on a range of other margins. We find no evidence that POs learned over time that the treatments were effective (table E.3) .[25] The lower prices paid for the same items might be expected to lead to increases in the quantities demanded, but table E.5 shows that this was not the case, possibly because the end users

[24]We can also control for composition effects arising from office-level variation in the types of items demanded by using office fixed effects ina difference in differences analysis that incorporate year 1 of the project. Appendix table E.2 shows that the treatment effects are, if anything, slightly larger.

[25]While we do not find evidence of experience effects, we *do*, however, find that the experiment had larger effects on offices for whom generic goods form a larger share of their annual budget, as shown in table E.4. These are offices where purchasing generics is a larger part of the job of the procurement officer and so the treatments have a bigger impact as one would expect.

for whom the items are being procured have inelastic demand. Finally, we see no evidence that procurement officers change the timing of their procurement to exploit predictable seasonality in prices or to attempt to game the incentives treatment (figure E.1).[26]

Overall, we conclude that on average, providing procurement officers with additional autonomy led to reduced prices without having an effect on the variety of goods purchased, the amount or composition of goods purchased, or the timing of procurement expenditure. We also do not see evidence of strong effects of the incentives treatment on any outcome.

To benchmark these findings, figure 4 shows a cost benefit evaluation of the implied savings. Savings are calculated as $\frac{-\eta_k}{1+\eta_k} \sum_o \text{Expenditure}_o \times \text{Treatment}_o^k$ where η_k are the estimated treatment effects in table 2 and Expenditure_o is the total spending by office o on generic goods (standard errors are calculated by the delta method). The solid lines denote savings net of the cost of the incentives treatment, while dashed lines are gross savings.

The figure reinforces our findings. The incentives treatment led to modest savings, while the autonomy and combined treatments led to large savings. The point estimate of the savings from the autonomy treatment is larger than the upper bound of the 95% confidence interval on the net savings from the incentives treatment. For comparison, the figure also shows the cost of operating 150 hospital beds, and the cost of operating 10 schools. Our point estimates suggest that the savings from the autonomy treatment from the relatively small group of offices in our experiment are sufficient to fund the operation of an additional 5 schools or to add 75 hospital beds.

For the incentives and combined treatments, the figure also shows the implied rates of return on the performance pay bonus payments. Despite the modest savings from the incentives treatment, these calculations imply a 45% rate of return on the incentives treatment since the small per-purchase savings are applied to a large base of expenditure. This rate of return is comparable to what Khan *et al.* (2016) find for performance payments to property tax inspectors in the same context.

The model presented in section 4 can provide structure to our interpretation of these findings. Our findings are consistent with what our model in section predicts will happen when the average monitor is relatively misaligned (high θ_{AG}). In particular, proposition 1 suggests that when the monitor is relatively misaligned the treatment effect of autonomy will be sizeable, while proposition 2 suggests that the treatment effect of incentives will be modest. Moreover, the model also provides predictions on how the treatment ef-

[26] In our pre-analysis plan we also listed budget utilization (whether or not bureaucrats are successfully able to spend their entire budget before the end of the fiscal year) as an outcome of interest in order to study changes in demand. As detailed in the populated pre-analysis plan (Duflo *et al.* , 2020) (available at https://doi.org/10.1257/rct.610-5.0) we also did not find effects on this outcome.

fects will vary with Monitor and PO alignment that allow us to shed further light on the mechanisms at play, issues to which we now turn.

6 Mechanisms

6.1 Monitor Alignment

Our conceptual framework in section 4 shows that shifting authority to the agent lowers prices only when the incentives of the agents are better aligned than those of the monitor. It thus predicts that we should expect to see heterogeneity in the treatment effects according to the alignment of the Accountant General (AG), θ_{AG}. In particular, the model predicts that the beneficial effects of the autonomy treatment should be concentrated among POs monitored by a relatively misaligned AG (high θ_{AG}) while the effects of the incentives treatment should be seen when the AG is well aligned (low θ_{AG}). In this section we estimate heterogeneous treatment effects using a proxy for the alignment of the AG.[27]

Each district has its own AG office and so we construct a proxy for each district AG's misalignment that combines two elements. First, we note that the main power of the accountant general is to delay payments and require additional paperwork. Second, in Punjab, as is common around the world, government offices' budgets lapse at the end of the fiscal year if they remain unspent. As documented in Liebman & Mahoney (2017) in the US context, lapsing budgets lead to a rush to spend at the end of the year. Combined with the first element, we expect this end of year rush to be stronger in districts where the accountant general delays payments more. Our proxy for the misalignment of the accountant general monitoring an office $\hat{\theta}_{AG,o}$ is therefore the fraction of purchases in the district in year 1 that were approved in June, the last month of the fiscal year.[28]

We augment equation (2) to include interactions with our proxy $\hat{\theta}_{AG,o}$ semi-parametrically

[27] In our pre-analysis plan we did not pre-specify that we would study heterogeneity by the accountant general's type. As the experiment rolled out and we discussed its impacts with our study participants we came to realize the importance of the type of the accountant general in determining how the treatments, particularly the autonomy treatment, affected the way procurement officers were able to change the way they carried out procurement.

[28] Appendix figure A.10 shows that the variation in this measure is not driven by variation across districts in the rate at which POs submit bills at the end of the year. Even conditional on the share of bills submitted at the end of the year, there is significant variation in the share of bills *approved* at the end of the year. We measure the fraction of purchases approved in June in our POPS data. However, the results are robust to measuring this in the finance department's administrative data instead.

using the approach of Robinson (1988) as follows

$$p_{igto} = \beta v_{igto} + \rho_g q_{igto} + \delta_s + \gamma_g + f\left(\hat{\theta}_{AG,o}\right) + \sum_{k=1}^{3} \text{Treatment}_o^k \times t_k\left(\hat{\theta}_{AG,o}\right) + \varepsilon_{igto}$$

where terms are as previously defined, $f(\cdot)$ is a non-parametric function of AG misalignment, and $t_k(\cdot)$ are non-parametric treatment effect functions.[29] Figure 5 shows the results.

Three key findings emerge consistent with the predictions of the model. First, the incentives treatment does reduce prices when the monitor is relatively more aligned (low $\hat{\theta}_{AG,o}$), and the treatment effect of incentives shrinks to zero as monitors become less aligned.[30] The treatment effect reaches zero when the June approval share is 0.48. Second, the autonomy treatment reduces prices more strongly when the monitor is relatively misaligned, with the treatment effect shrinking to zero when the June share drops below 0.22. The top of the figure shows the coefficients, standard errors, and p-values from difference-in-differences regressions using these thresholds to classify good and bad AGs, and we use these thresholds in our analysis going forward.[31] Third, the broad range of AG misalignment over which the autonomy treatment is effective suggests that the competing bandits effect highlighted in corollary 1 is at play. We expand upon this below in section 6.2. Overall, the results are remarkably consistent with the predictions of the model, and suggest that the average effects of the treatments are more consistent with the average AG being relatively misaligned.[32]

[29] To implement this we rewrite the model as $p_{igto} = \mathbf{x}_{igto}\beta + f(\hat{\theta}_{AG,o}) + \sum_{k=1}^{3} \text{Treatment}_o^k \times t_k(\hat{\theta}_{AG,o}) + \varepsilon_{igto}$ and proceed in four steps. First, we run treatment-group specific non-parametric regressions of p_{igto} on $\hat{\theta}_{AG,o}$ to form conditional expectations $\mathsf{E}[p_{igto}|\hat{\theta}_{AG,o}, \text{Treatment}_o^k] \simeq \hat{m}_k(\hat{\theta})$ and linear regressions of the control variables $\mathbf{x}_{igto} = \alpha + \xi\hat{\theta}_{AG,o} + \sum_{k=1}^{3}(\eta_k \text{Treatment}_o^k + \zeta_k \text{Treatment}_o^k \times \hat{\theta}_{AG,o}) + \varepsilon_{igto}$ to form conditional expectations $\mathsf{E}[\mathbf{x}_{igto}|\hat{\theta}_{AG,o}, \text{Treatment}_o^k] \simeq \hat{j}(\hat{\theta})$. Second, we regress $p_{igto} - \hat{m}_k(\hat{\theta}) = [\mathbf{x}_{igto} - \hat{j}(\hat{\theta})]\beta + \varepsilon_{igto}$. Third, we non-parametrically regress $p_{igto} - \mathbf{x}_{igto}\hat{\beta} = r_k(\hat{\theta}_{AG,o}) + \varepsilon_{igto}$ separately in the control group ($k = 0$) and the three treatment groups. Fourth, we form the estimates $\hat{f}(\hat{\theta}_{AG,o}) = \hat{r}_0(\hat{\theta}_{AG,o})$ and $\hat{t}_k(\hat{\theta}_{AG,o}) = \hat{r}_k(\hat{\theta}_{AG,o}) - \hat{r}_0(\hat{\theta}_{AG,o}),\ k = 1, \ldots, 3$.

[30] It should be noted, though, that since each district has its own AG and they are not randomly assigned, these results should be interpreted with caution. If our AG measure is correlated with other district-level factors that drive the impacts of the experiment, we may be picking those up instead of the effect of AG misalignment. Reassuringly, our AG misalignment measure is not correlated with a range of district-level observable features including population, area, health, education and poverty indices suggesting at least that these are not the factors driving the results.

[31] Appendix figure F.1 and table F.1 use simple linear difference in differences specifications to explore the sharp jumps revealed by the nonparametric analysis and show its robustness to our alternative ways of controlling for item variety.

[32] Online Appendix F shows the robustness of these findings for prices, and that there is no heterogeneity in the effects on other outcomes. Table F.2 shows that the effects are robust to using three alternative proxies for AG type. Table F.3 shows robustness to three potential confounders (delayed submission by POs, average delays, and average PO type) our proxy for AG type may be picking up. Tables F.4 and F.5 show that the effects on item variety, and quantity demande, are not heterogeneous, respectively.

Figure 6 extends figure 4 to show how the cost benefit calculus of the treatments varies with the degree of misalignment of the AG. The vertical axis measures for each district the total net savings by all districts with a less misaligned accountant general: $\sum_{d:j_d \leq x} \left[\left(\frac{-\eta_k(j_d)}{1+\eta_k(j_d)} \sum_{o \in d} \text{Expenditure}_{od} \times \text{Treatment}_o^k \right) - c_d \right]$ where $\eta_k(j_d)$ are estimated treatment effects of treatment k when monitor misalignment is j_d and c_d is the ex ante cost of performance pay bonuses to offices in district d (the number of offices in the district at each pay grade times the expected prize for each office). The figure shows large net savings from the incentives treatment, even at low levels of misalignment. By contrast, net savings from the autonomy and combined treatments are negligible in districts with low misalignment; they only accrue at high levels of monitor misalignment.

To better understand how the misalignment of the monitor matters for prices, we analyze the effects of the treatments on the main power that the AG has in the status quo—to delay and hold up approval of purchases. We study delays between a purchase and its approval by the AG through a series of distributional regressions of the probability of delay of at least j days in year 2 normalized by the probability of a delay of at least j days in the control group in year 1 on treatment dummies, strata fixed effects γ_s and good fixed effects γ_g:

$$\frac{\mathbf{1}\left\{\text{delay}_{igo} \geq j\right\}}{\mathbb{P}\left(\text{delay} \geq j | \text{Control, Year1}\right)} = \alpha + \sum_{k=1}^{3} \eta_k \text{Treatment}_o^k + \gamma_s + \gamma_g + \varepsilon_{igo}$$

Panel A of figure 7 shows the results, and also shows the CDF of delays in the control group in year 1 for reference. We clearly see a decrease in very long delays in the autonomy treatment, and very little effect in the other treatments. Panel B separates the effect of the autonomy treatment for good (June share ≤ 0.22) and bad (June share > 0.22) AGs, showing that the effect on long, costly delays is driven exclusively by offices facing a more misaligned monitor while POs facing a good AG only see reductions in shorter delays.

Since vendors have to make deliveries before being paid, these delays are costly to both the vendors and to the POs and one would naturally expect vendors to charge POs a markup for the delays. When POs in the autonomy group can pay vendors immediately in cash, the removal of these markups may contribute to the effect of the autonomy treatment on prices. However, note that the removal of these markups cannot fully account for the estimated treatment effect of autonomy. Even assuming that the petty cash allows POs to completely avoid delays of six months would require that vendors charge interest of 242% to account for the price savings, far above market interest rates.[33]

[33] To see this, note that a PO with a budget B who faces an interest rate r and a delay of t years to

Nevertheless, this effect on overall delays could be driven by general inefficiency of the AG or by POs dragging their feet in submitting paperwork. We therefore focus on delays that are more clearly suggestive of holdup: purchases that are approved right at the end of the fiscal year. We analyze how the treatments change the probability that items purchased in different months are approved in June (the last month of the fiscal year) by estimating equations of the form

$$\begin{aligned}\mathbf{1}\left\{\text{Approved in June}_{igo}\right\} = \alpha + \sum_{k=1}^{3}\sum_{m=Jul}^{Jun}\eta_{mk}\mathbf{1}\left\{\text{PurchaseMonth}_{igo} = m\right\} \times \text{Treatment}_o^k \\ + \sum_{m=Jul}^{Jun}\gamma_m\mathbf{1}\left\{\text{PurchaseMonth}_{igo} = m\right\} + \gamma_g + \varepsilon_{igo}\end{aligned}$$

Panel A of figure 8 shows the η_{mk} coefficients for the autonomy treatment and also the raw distribution of delivery dates of purchases approved in June in the autonomy treatment (in orange) and control (in green) groups. It clearly shows that purchases at the beginning of the year (in July and August in particular) are much less likely to have to wait right until the end of the year to be approved, strongly suggesting that the holdup power of the AG has been decreased. Panel B runs the regression separately for good (June share ≤ 0.22) and bad (June share > 0.22) AGs, and shows that this reduction in holdup for purchases made at the beginning of the year is exclusively driven by the bad AGs. Overall, the results suggest that monitor misalignment is a key driver of the effects of the experimental treatments, and that it affects prices through the ability of the AG to hold up purchases.

6.2 Procurement Officer Alignment

Our conceptual framework in section 4 suggests that the impacts of the experiment will be heterogeneous by the degree of misalignment θ_{PO} of the procurement officer. At baseline, we collected one potential proxy for the PO's type—the lab-in-the-field measure of dishonesty studied in Fischbacher & Föllmi-Heusi (2013) and Hanna & Wang (2017). However, as shown in appendix figure A.11, the POs' scores are not predictive of prices at baseline, suggesting these scores are not successfully capturing POs' types.[34] Unsurprisingly, as table A.4 shows, the dice scores also do not predict heterogeneity in the treatment

pay vendors can spend a pre-markup amount of $S = B(1+r)^{-t}$. If that PO has the same spending but can make 100K worth of spending in cash, then their total spending would be $B\left(1-\eta_{\text{Autonomy}}\right) = (S-100K)(1+r)^t + 100K$. To account for a saving of $\eta_{\text{Autonomy}} = 0.085$ when the PO has the average budget of 1 million Rupees and the delay is $t = 0.5$ years requires an annual interest rate of 242%.

[34] Despite there being significant variation across POs in their dice scores (as shown in panel A of figure A.11)

effects. Our pre-analysis plan also listed a range of officer-level (Tenure, pay scale, education) and office-level (distance from the AG, distance from the department HQ) traits that might predict heterogeneity. As detailed in the populated pre-analysis plan, (Duflo *et al.* , 2020) these also did not predict heterogeneity in the treatment effects.[35]

The lack of heterogeneity by this proxy for Procurement Officer (PO) type does not, however, suggest that the POs in our setting are all well aligned. We can see this by viewing our results on the heterogeneity of the treatment effects by AG alignment through the lens of our model in section 4. Figure 5 reveals three distinct regions. For highly aligned AGs, the incentives treatment reduces prices but the autonomy treatment does not. For intermediate values of θ_{AG} both treatments are effective, and for high values of θ_{AG} the autonomy treatment reduces prices but the incentives treatment does not. Comparing this to the theoretical predictions in propositions 1 and 2 and corollary 1 (summarized in figure 3), these results are consistent with the model when the average PO is relatively misaligned (high θ_{PO}), but not what we would expect if the POs are well aligned (low θ_{PO}). With misaligned POs, incentives reduce prices under relatively well aligned AGs (2), and due to the competing bandit effect autonomy reduces prices even for intermediate levels of AG alignment by eliminating double marginalization (corollary 1).

There is also direct evidence that POs are relatively misaligned. Our endline survey asked POs about a range of potential mechanisms, focusing on how much time POs and their staff spend on procurement and how they allocate their time across different procurement tasks. Figure 9 shows that all three treatments increase the amount of time POs report spending on procurement. POs in the autonomy treatment increase the time they spend on procurement by 16%. Similarly, POs in the incentives treatment increase the time they spend on procurement by 14%, and those in the combined treatment by 20%.[36] If POs were already well aligned, we would not expect them to increase their time allocation in response to the treatments. Moreover, appendix figure A.12 shows how the treatments change the way that POs allocate time across different tasks. Panel A shows that POs in the autonomy and combined treatments spend less time instructing their staff and negotiating approvals with the AG, consistent with the autonomy treatment restricting the AG's holdup power over the POs. Panel B shows that POs in the autonomy treatment are less interested in choosing vendors who are able to provide them credit or help them negotiate approvals at the AG. POs thus spend less time themselves on dealing with the AG and are able to focus on vendors who provide better goods rather than vendors who are able to help POs navigate the AG's office.

[35]The populated pre-analysis plan is available at https://doi.org/10.1257/rct.610-5.0.

[36]We cannot reject the hypothesis that the three increases are the same ($p = 0.70$).

Changes in time use by POs are also able to explain the majority of the effects of the treatments on prices. Table 3 quantifies how much of the estimated effect comes from changes in behavior of the PO by estimating the effect of treatment on prices via their effect on time devoted to procurement. The first two columns of Table 3 test whether in these cases treatment affects time spent on procurement when we know it to affect prices.[37] Note that, in theory, the effect of autonomy on time spent on procurement is ambiguous because treatment decreases the time needed to deal with the AG but it might increase the time POs devotes to finding good deals now that they do not have to wait for the AG. Column 1 shows that the first effect prevails: when the AG is bad, giving autonomy to the PO increases time spent by 20% . In theory, incentives increase effort and, in line with this, Column 2 shows that when the AG is good, providing incentives increases time spent on procurement by 24%.

Having established that the treatments affect time devoted to procurement, we can use this as a first stage to quantify the effect of treatment on prices. Back of the envelope calculation indicates that the IV estimate of the effect of time on prices, -0.012 combined with the first-stage estimate of the increase in time spent, 6.98, suggests that increases in time spent account for a price decrease of $6.98 \times 0.012 = 0.084$ log points, 62% of the price decrease of 0.14 log points when the AG is bad estimated in table F.1. Similarly, when the AG is *good,* the estimates in columns 4–6 suggest that increases in time spent in the incentives treatment account for a price decrease of $8.3 \times 0.010 = 0.083$ log points, 72% of the 0.12 log point price decrease estimated in table F.1. To be clear, these capture the effect of all the PO's actions that are correlated with the time he devotes to procurement. They are not, in other words, causal impacts of time on prices.

Finally, columns 5-6 report, as a placebo, the two cases where we know the treatment had no effect, that is incentives with a bad AG and autonomy with a good AG. In line with the earlier results, we see that the treatments do not affect the time devoted to procurement. In other words, when treatment does not affect prices it does not affect time either. As we have no first stage we do not report second stage estimates in this case. Overall, the experimental treatments induced POs to devote more effort to procurement, reducing prices whenever POs faced a monitoring environment conducive to improved performance, but not when the monitoring environment prevents POs from improving performance, consistent with the theoretical framework in section 4.

[37] We use the share of June approvals measure discussed in section 6.1. A good AG has a June share of 0.22 or below in panels A and C, and 0.48 or below in panel B

7 Conclusion

Recent advances in the empirical analysis of organizations have improved our understanding of the relationship between principals and agents and how management practices such as performance pay and decentralization shape organizations' performance. Most organizations, however, are more complex than the single-layer theoretical construct we use to analyze them. Control over rules and incentives that regulate agents' behavior resides with other agents at higher levels of the hierarchy rather than with the principal herself, and these agents might also be prone to act in their own interest.

Our experiment shows that the allocation of authority between agents at different levels of the hierarchy shapes the performance of the organization, and that this depends on the relative severity of misalignment of different agents. Similarly, the effect of providing incentives on performance also depends on how authority is allocated between agents. Hence, the two must be designed jointly to ensure compatibility. Shifting authority to frontline agents reduces the prices the bureaucracy pays for its inputs by 9% on average, and up to 15% when the monitor is more inefficient or corrupt. The mechanism through which this happens is the reduction of long delays in monitor approvals. This increases taxpayers' welfare at the expense of the monitors' and possibly also sellers' who were charging higher prices for longer waits.

The monitors and the monitored tend to come from the same culture, face the same institutional incentives and be exposed to the same temptations. In these circumstances, adding a monitor with veto power is a bad idea. If a country sees high levels of corruption, it is a natural reaction to call for more monitoring but it can do more harm than good as we now have two bandits instead of one. To do better, we must design more sophisticated institutions that are more robust to misaligned agents. For instance, ex post monitoring is less manipulable as the PO has less incentive to bribe a monitor who cannot veto the purchase. The monitor can then be financially motivated to impose discipline by being promised a share of aggregate savings.

The results raise several questions for future research. First, if rules are so costly why do most bureaucracies use them? One possibility is that large corruption "scandals" are much more damaging to the organization than the, potentially much larger, sum of small markups on a large volume of smaller transactions. Our benchmarking exercise suggests that the cost created by corruption scandals must exceed 10 million rupees for the stringent rules to be a rational choice. Figure 10 provides evidence on whether such scandals, that is extremely high prices, are common in our treatment groups. The figure reports quantile treatment effect estimates. If autonomy made scandals more likely, we would expect to see that the 9% average reduction was masking large increases in prices at the high quantiles of the price distribution. If anything, we see the opposite: the treatment effects of all three treatments are negative at the higher quantiles.

Second, we have studied the effect of shifting authority in an organization while keeping the selection of agents into the organization constant. It is well-known that different incentives attract different types of workers (Dal Bó *et al.* , 2013; Ashraf *et al.* , 2020; Deserrano, 2019), for instance performance pay typically attracts workers with better skills who can benefit from performance rewards (Lazear, 2000). In our case, more autonomy might attract officers who are more prone to exploit it to their personal advantage. At the same time, giving more autonomy to officers implies taking it away from the monitors and therefore the treatment might attract monitors who are less likely to exploit their position for private gain.

Finally, we have focused on the procurement of homogeneous goods. While these goods represent over half the procurement done by the offices in our sample, we leave unanswered the question of how autonomy might affect the procurement of more complex goods. For example, while in our setting we do not see responses along these margins, for more complex products the quantities procured (Lichand & Fernandes, 2019) and the quality of the item procured may respond as limits to the ability to contract on quality become first order (Bosio *et al.* , 2020).

Our results have implications for the design and interpretation of field experiments within organizations. It is very common for researchers to effectively become the principal during the implementation of different policies in order to achieve control. This is innocuous to the extent that they share the same objectives if not the same skills. However it is not innocuous if researchers are not replacing the principal, but rather another set of agents who have different incentives. This has implications for the scalability of the results and can explain why interventions which are very successful when implemented by researchers do not work when implementation is delegated to managers or other agents.[38]

[38] Examples include the "camera" experiment by Duflo *et al.* (2012) that was successfully implemented by researchers but failed when implemented by the government, because staff who were supposed to enforce punishments failed to do so (Banerjee *et al.* , 2008). Similarly, incentive contracts offered to teachers in Kenya by an international NGO were effective whilst the same contracts failed when monitored by the government (Bold *et al.* , 2018)

References

Ashraf, Nava, Bandiera, Oriana, Davenport, Edward, & Lee, Scott S. 2020. Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services. *American Economic Review*, **110**, 1355–1394.

Bajari, Patrick, & Tadelis, Steven. 2001. Incentives vs Transaction Costs: A Theory of Procurement Contracts. *RAND Journal of Economics*, **32**, 387–407.

Bandiera, Oriana, Prat, Andrea, & Valletti, Tommaso. 2009. Active and Passive Waste in Government Spending: Evidence from a Policy Experiment. *American Economic Review*, **99**(4), 1278–1308.

Bandiera, Oriana, Barankay, Iwan, & Rasul, Imran. 2011. Field Experiments with Firms. *Journal of Economic Perspectivs*, **25**, 63–82.

Banerjee, A, Hanna, R, & Mullainathan, S. 2012. Corruption. *Chap. 27, pages 1109–1147 of: Handbook of Organizational Economics*.

Banerjee, Abhijit V., Duflo, Esther, & Glennerster, Rachel. 2008. Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System. *Journal of the European Economic Association*, **6**, 487–500.

Best, Michael Carlos, Hjort, Jonas, & Szakonyi, David. 2019. *Individuals and Organizations as Sources of State Effectiveness*. Mimeo: Columbia University.

Bloom, Nicholas, Lemos, Renata, Sadun, Raffaela, & Van Reenen, John. 2015a. Does Management Matter in Schools? *Economic Journal*, **125**, 647–674.

Bloom, Nicholas, Propper, Carol, Seiler, Stephan, & Van Reenen, John. 2015b. The Impact of Competition on Management Quality: Evidence from Public Hospitals. *Review of Economic Studies*, **82**, 457–489.

Bold, Tessa, Kimenyi, Mwangi, Mwabu, Germano, Ng'ang'a, Alice, & Sandefur, Justin. 2018. Experimental Evidence on Scaling up Education Reforms in Kenya. *Journal of Public Economics*, **168**, 1–20.

Bolton, Patrick, & Dewatripont, Mathias. 2013. Authority in Organizations: A Survey. *Chap. 9, pages 342–372 of:* Gibbons, Robert S., & Roberts, John (eds), *The Handbook of Organizational Economics*. Princeton University Press.

BOSIO, ERICA, DJANKOV, SIMEON, GLAESER, EDWARD, & SHLEIFER, ANDREI. 2020. *Public Procurement in Law and Practice*. Mimeo: Harvard University.

COASE, RONALD. 1937. The Nature of the Firm. *Economica*, **4**, 386–405.

COVIELLO, DECIO, GUGLIELMO, ANDREA, & SPAGNOLO, GIANCARLO. 2018. The Effect of Discretion on Procurement Performance. *Management Science*, **64**, 715–738.

DAL BÓ, ERNESTO, FINAN, FEDERICO, & ROSSI, MARTIN A. 2013. Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service. *Quarterly Journal of Economics*, **128**, 1169–1218.

DESERRANO, ERIKA. 2019. Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda. *American Economic Journal: Applied Economics*, **11**, 277–317.

DUFLO, ESTHER, HANNA, REMA, & RYAN, STEPHEN P. 2012. Incentives Work: Getting Teachers to Come to School. *American Economic Review*, **102**, 1241–1278.

DUFLO, ESTHER, GREENSTONE, MICHAEL, PANDE, ROHINI, & RYAN, NICHOLAS. 2018. The Value of Discretion in the Enforcement of Regulation: Experimental Evidence and Structural Estimates from Environmental Inspections in India. *Econometrica*, **86**, 2123–2160.

DUFLO, ESTHER, BANERJEE, ABHIJIT, FINKELSTEIN, AMY, KATZ, LAWRENCE F., OLKEN, BENJAMIN A., & SAUTMANN, ANJA. 2020. *In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics*. NBER Working Paper No. 26993.

DUGGAN, MARK, & SCOTT MORTON, FIONA M. 2006. The Distortionary Effects of Government Procurement: Evidence from Medicaid Prescription Drug Purchasing. *Quarterly Journal of Economics*, **121**, 1–30.

FINAN, FRED, OLKEN, BEN, & PANDE, ROHINI. 2017. The Personnel Economics of the Developing State. *Chap. 6, pages 467–514 of:* BANERJEE, ABHIJIT, & DUFLO, ESTHER (eds), *Handbook of Economic Field Experiments*, vol. 2. Elsevier.

FISCHBACHER, URS, & FÖLLMI-HEUSI, FRANZISKA. 2013. Lies in Disguise: an Experimental Study on Cheating. *Journal of the European Economic Association*, **11**, 525–547.

GERARDINO, MARIA, LITSCHIG, STEPHAN, & POMERANZ, DINA. 2017 (10). *Distortion by Audit: Evidence from Public Procurement*. Mimeo: University of Zurich.

GIBBONS, ROBERT, & ROBERTS, JOHN. 2013. Economic Theories of Incentives in Organizations. *Chap. 2, pages 100–158 of:* GIBBONS, ROBERT, & ROBERTS, JOHN (eds), *Handbook of Organizational Economics*. Princeton University Press.

GURIEV, SERGEI. 2004. Red tape and corruption. *Journal of Development Economics*, **73**(2), 489–504.

HANNA, REMA, & WANG, SHING-YI. 2017. Dishonesty and Selection into Public Service: Evidence from India. *American Economic Journal: Economic Policy*, **9**(3), 262–90.

IMBENS, GUIDO W., & RUBIN, DONALD B. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

KELMAN, STEVE. 1990. *Procurement and Public Management: The Fear of Discretion and the Quality of Government Performance*. AEI Press.

KHAN, ADNAN Q., KHWAJA, ASIM I., & OLKEN, BENJAMIN A. 2016. Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors. *Quarterly Journal of Economics*, **131**, 219–271.

LAFFONT, JEAN-JACQUES, & TIROLE, JEAN. 1994. *A Theory of Incentives in Procurement and Regulation*. MIT Press.

LAZEAR, EDWARD. 2000. Performance Pay and Productivity. *American Economic Review*, **95**, 1346–1361.

LEAVER, CLARE. 2009. Bureaucratic Minimal Squawk Behavior: Theory and Evidence from Regulatory Agencies. *American Economic Review*, **99**(05), 572–607.

LICHAND, GUILHERME, & FERNANDES, GUSTAVO. 2019. *The Dark Side of the Contract: Do Government Audits Reduce Corruption in the Presence of Displacement by Vendors?* Tech. rept. Working Paper.

LIEBMAN, JEFFREY B., & MAHONEY, NEALE. 2017. Do Expiring Budgets Lead to Wasteful Year-End Spending? Evidence from Federal Procurement. *American Economic Review*, **107**, 3510–3549.

NATIONAL AUDIT OFFICE. 2006. *Assessing the value for money of OGCbuying.solutions*. https://www.nao.org.uk/wp-content/uploads/2006/12/0607103.pdf.

OECD. 2009. *OECD Principles for Integrity in Public Procurement*. Tech. rept. Organisation for Economic Co-operation and Development.

OLKEN, BENJAMIN A., & BARRON, PATRICK. 2009. The Simple Economics of Extortion: Evidence from Trucking in Aceh. **117**, 417–452.

PUNJAB PROCUREMENT REGULATORY AUTHORITY. 2014. *Punjab Procurement Rules No.ADMN(PPRA) 10–2/2013*. https://ppra.punjab.gov.pk/system/files/Final%20Notified%20PPR-2014%20%28ammended%20upto%2006.01.2016%29.pdf.

RASUL, IMRAN, & ROGGER, DANIEL. 2018. Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service. *Economic Journal*, **128**, 413–446.

RASUL, IMRAN, ROGGER, DANIEL, & WILLIAMS, MARTIN J. 2019. *Management, Organizational Performance, and Task Clarity:Evidence from Ghana's Civil Service*. forthcoming, Journal of Public Administration Research and Theory.

ROBINSON, PETER. 1988. Root-N-Consistent Semiparametric Regression. *Econometrica*, **56**, 931–954.

RUBIN, DONALD B. 1980. Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test" by Basu. *The Journal of the American Statistical Association*, **75**, 591–593.

SANCHEZ DE LA SIERRA, RAUL, & TITECA, KRISTOF. 2019. *Corruption (with a Hiearchy)*. Mimeo: University of Chicago.

SHI, LAN. 2008. The limit of oversight in policing: Evidence from the 2001 Cincinnati riot. *Journal of Public Economics*, **93**(07), 99–113.

SHLEIFER, ANDREI, & VISHNY, ROBERT W. 1993. Corruption. *Quarterly Journal of Economics*, **108**, 599–617.

SIMON, HERBERT. 1951. A Formal Theory of the Employment Relationship. *Econometrica*, **19**, 293–305.

SZUCS, FERENC. 2017. *Discretion and Corruption in Public Procurement*. Mimeo: Stockholm University.

WEAVER, JEFFREY. 2020. *Jobs for Sale: Corruption and Misallocation in Hiring*. Forthcoming, American Economic Review.

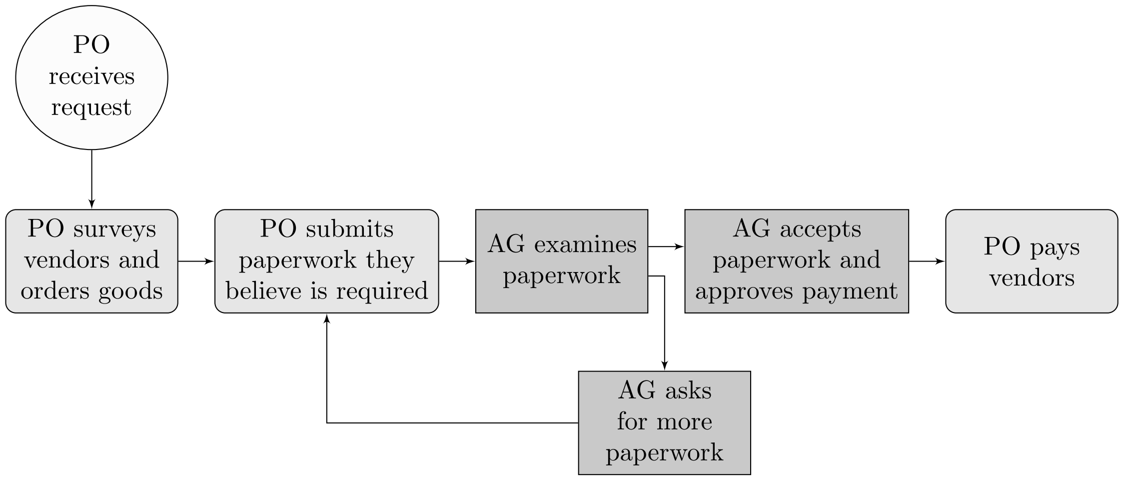
YANG, DEAN. 2008. Can Enforcement Backfire? Crime Displacement in the Context of Customs Reform in the Philippines. *The Review of Economics and Statistics*, **90**(1), 1–14.

YOUNG, ALWYN. 2019. Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *Quarterly Journal of Economics*, **134**, 557–598.

Figures & Tables

FIGURE 1: PROCUREMENT PROCESS SUMMARY

Panel A: Status Quo Procurement Process



Panel B: Procurement Process Under Autonomy Treatment

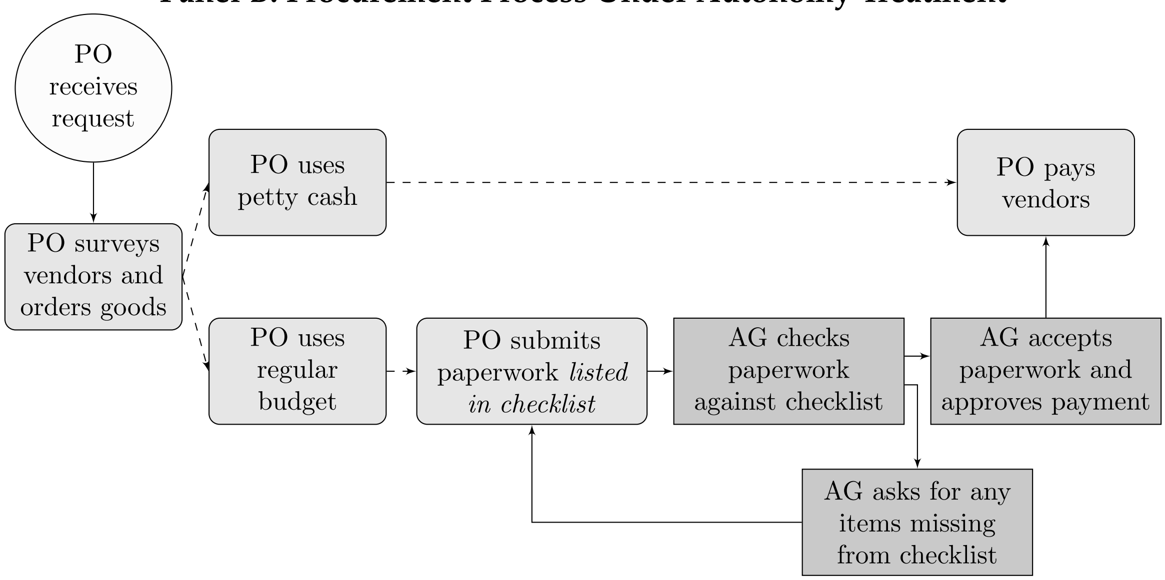
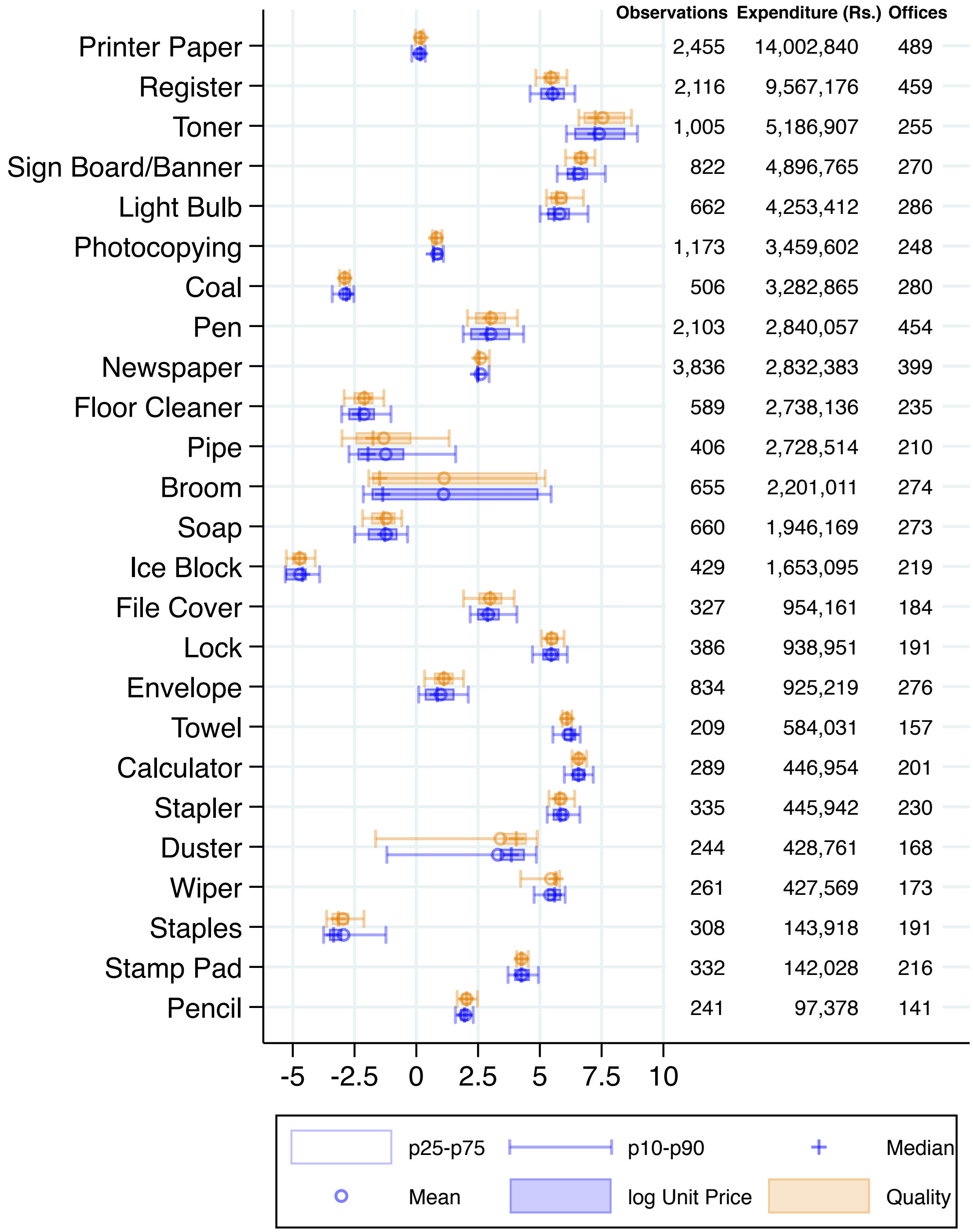
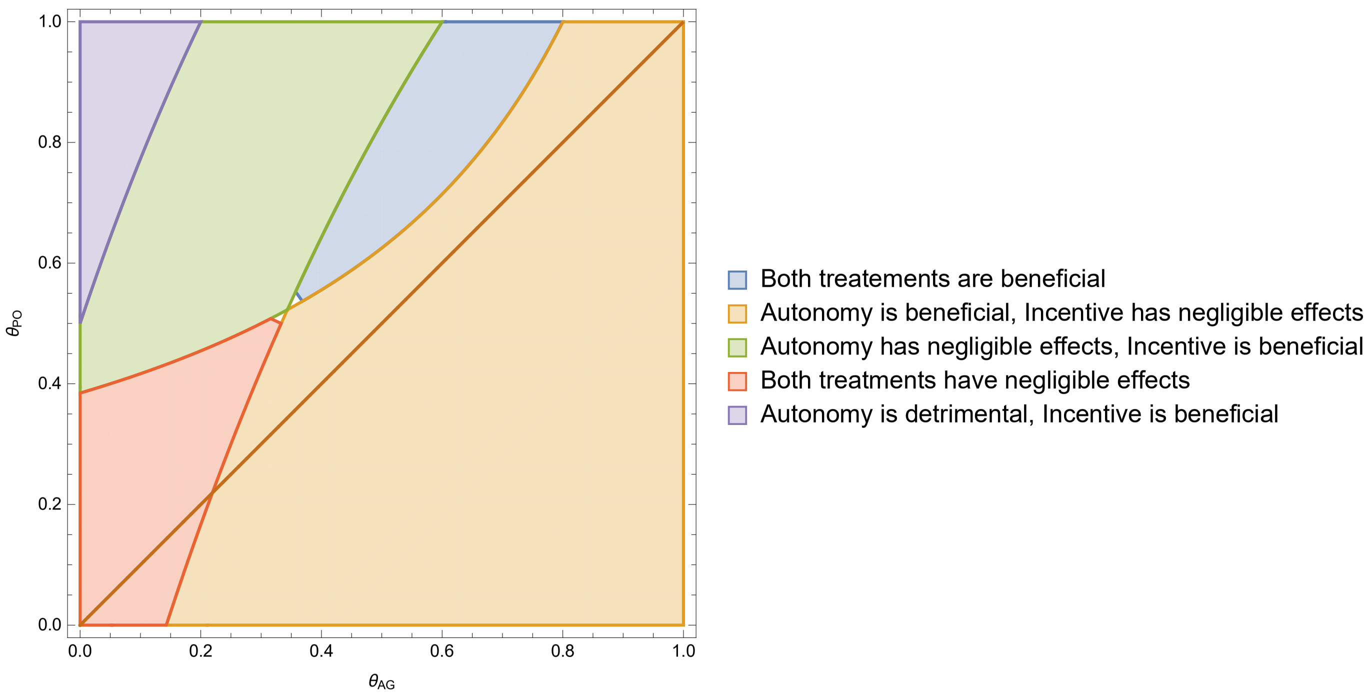


FIGURE 2: SUMMARY STATISTICS



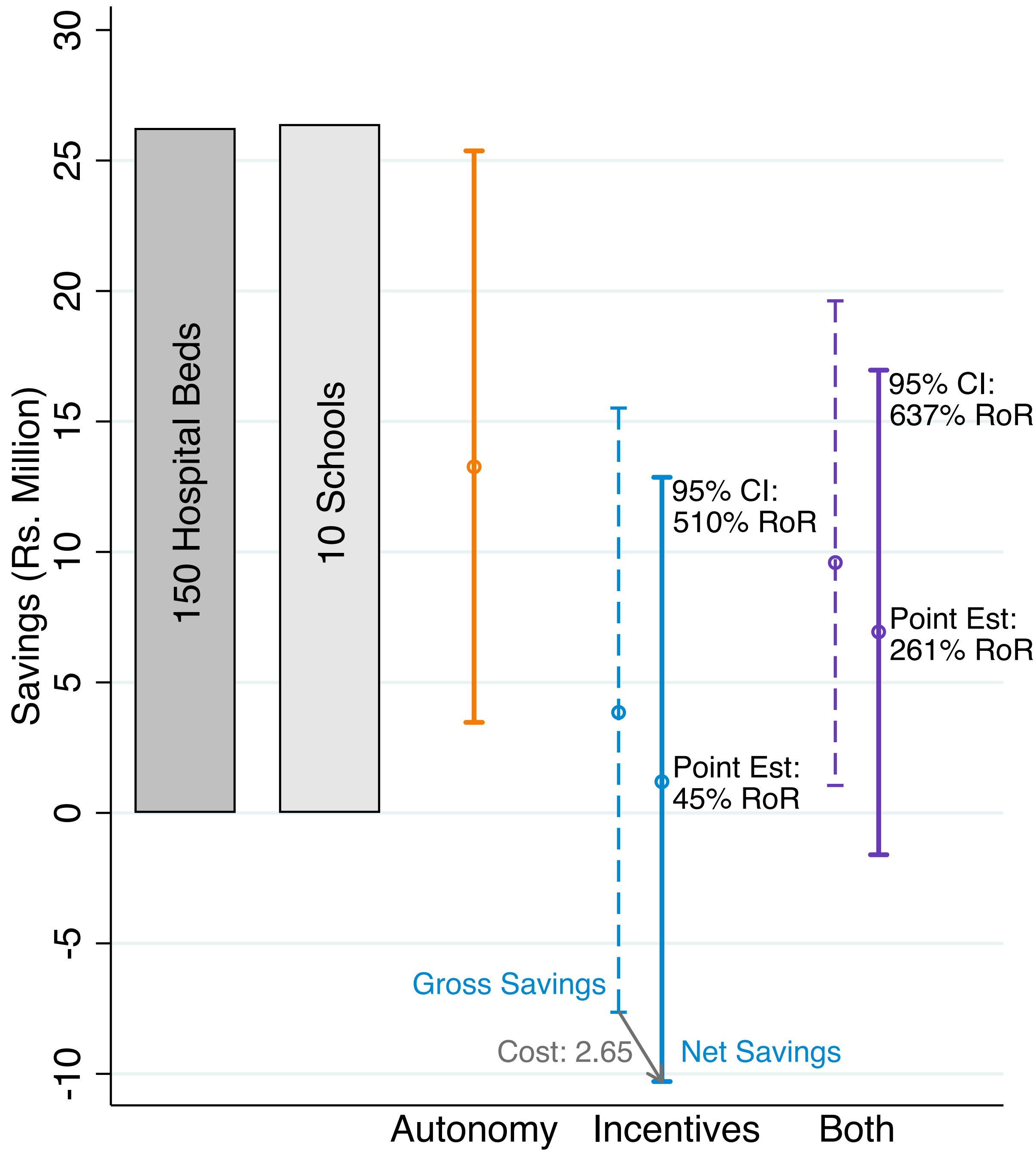
Notes: The figure displays summary statistics for the purchases of the goods in our cleaned purchase sample. The figure summarizes the log unit prices paid for the goods, the number of purchases of each good, and the total expenditure on the good (in Rupees) in the sample.

FIGURE 3: MODEL PREDICTIONS OF TREATMENT EFFECTS



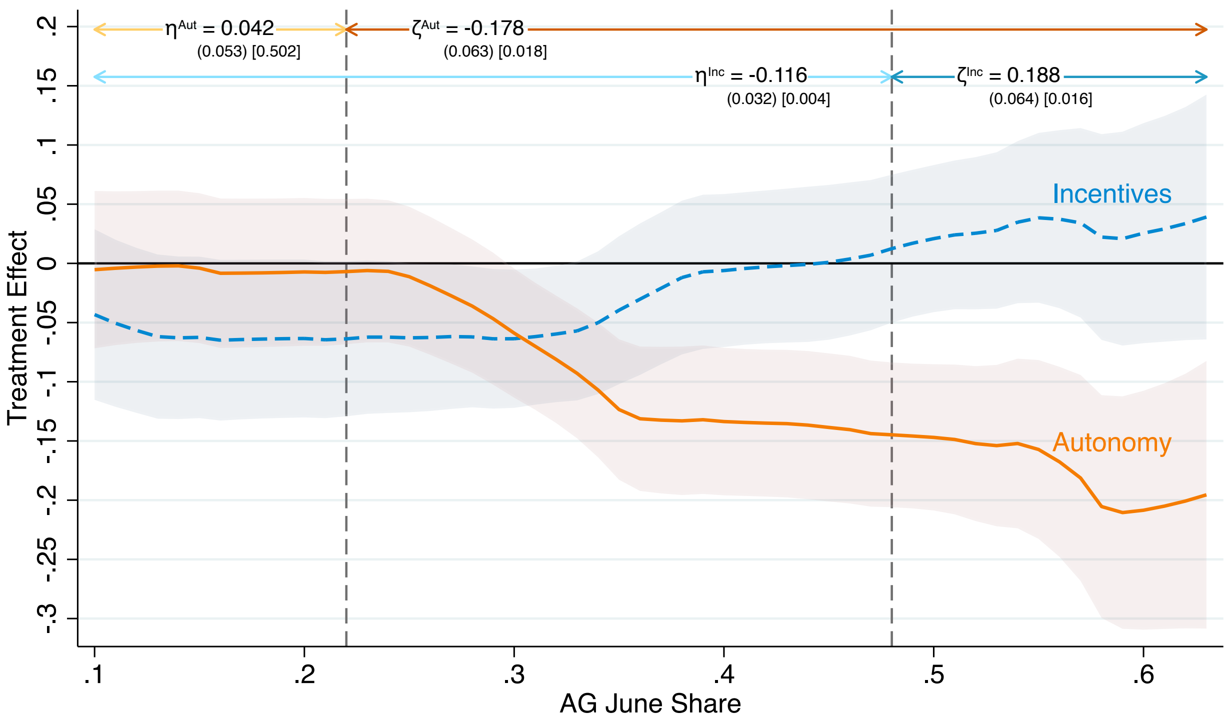
Notes: The figure shows the predictions our model in section 4 makes about how the treatment effects of our experiment will vary with the degree of misalignment of the monitor (θ_{AG}) and the degree of misalignment of the procurement officer (θ_{PO}) as described in proposition 1, corollary 1 and proposition 2. The plot is drawn for values of the parameters that satisfy the four assumptions above ($p_A = p_{AA} = c = 100$, $p_{MM} = 138$, $p_{AM} = 135$, $p_{MA} = 113$, $p_M = 123$). A treatment is "beneficial" if the expected price change is a reduction of at least 5% of the minimal price c. Conversely, a treatment is "detrimental" if it induces a price increase of at least 5%. A treatment has a "negligible effect" if the average price change is between -5% and +5%.

FIGURE 4: COST BENEFIT ANALYSIS



Notes: The figure shows a cost benefit analysis of the experiment. For each treatment, the vertical intervals denote total savings due to the experiment in millions of Rupees. Savings are calculated as $\frac{-\eta_k}{1+\eta_k} \sum_o \text{Expenditure}_o \times \text{Treatment}_o^k$ where η_k are the estimated treatment effects in table 2 and Expenditure_o is the total spending by office o on generic goods (standard errors are calculated by the delta method). The solid lines denote savings net of the cost of the incentives treatment, while dashed lines are gross savings. For the incentives and combined treatments, the figure also shows the implied rates of return on the performance pay bonus payments. For comparison, the figure also shows the cost of operating 150 hospital beds, and the cost of operating 10 schools.

FIGURE 5: HETEROGENEITY OF TREATMENT EFFECTS BY MONITOR ALIGNMENT

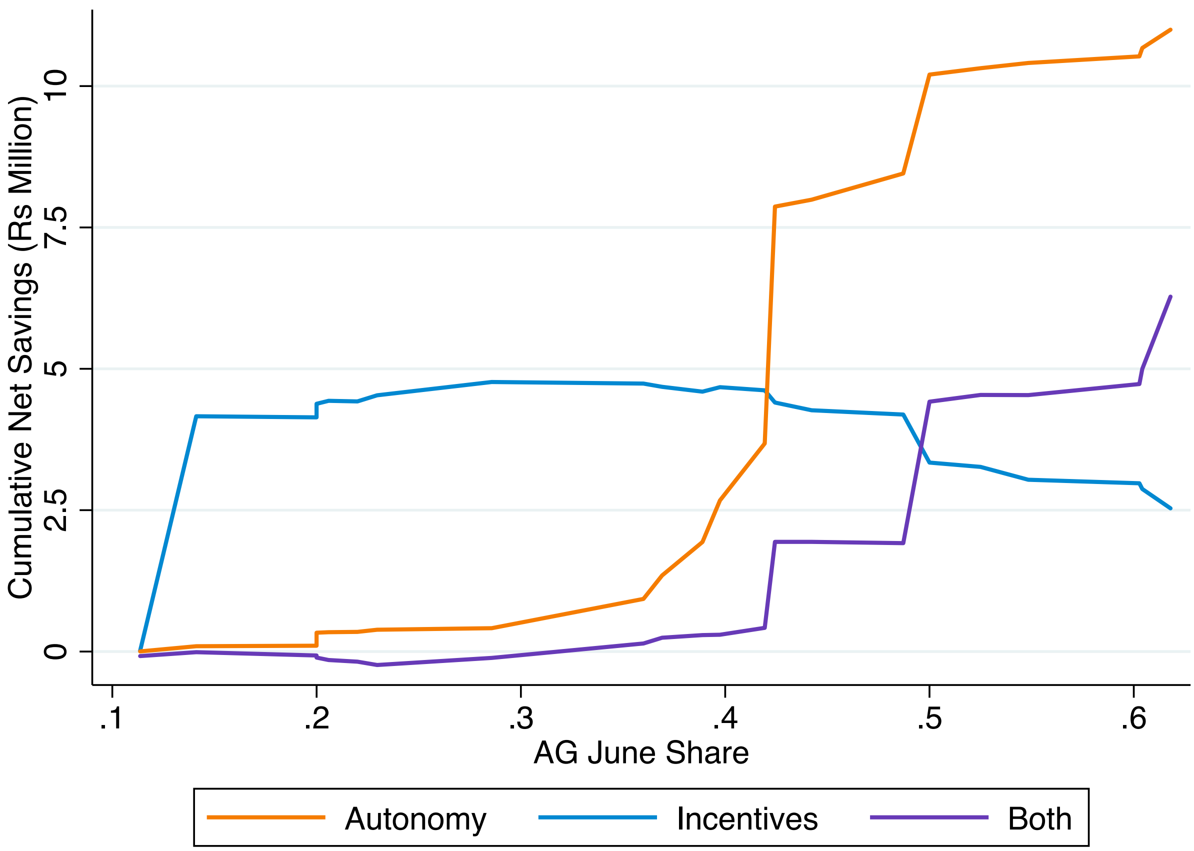


Notes: The figure shows heterogeneity of the treatment effects of autonomy and incentives by the degree of misalignment of the district's accountant general (AG). As discussed in section 6.1 AGs are classified according to the degree to which purchase approvals are bunched at the end of the fiscal year in June 2015 (year 1 of the project). The figure shows semi-parametric estimates of the treatment effects using the method in Robinson (1988) to estimate linear effects of the full set of controls and flexible non-parametric heterogeneous treatment effects by accountant general:

$$p_{igto} = \mathbf{X}_{igto}\beta + \sum_{k=1}^{3} f_k\left(\text{AGJuneShare}_o\right) \times \text{Treatment}_o^k + \varepsilon_{igto}$$

where $\mathbf{X}_{igto}$ includes the scalar item variety measure, good specific controls for purchase size, stratum FEs, and good fixed effects, and $f_k(\cdot)$ are nonparametric treatment effect functions. The top of the figure shows coefficients, clustered standard errors (in parentheses) and randomization inference p-values (in square brackets) from difference in differences regressions interacting treatment dummies with a dummy for facing a "bad" AG, decined as a June share of 0.22 for autonomy, and 0.48 for incentives. $p_{igto} = \alpha + \eta\text{Treatment}_o + \zeta\text{Treatment}_o \times \text{BadAG}_o + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$ where all terms are as defined above, and we control for the scalar measure of item variety as part of $\mathbf{X}_{igto}$. Appendix figure F.1 varies the thresholds used for defining a bad AG, justifying the use of 0.22 and 0.48.

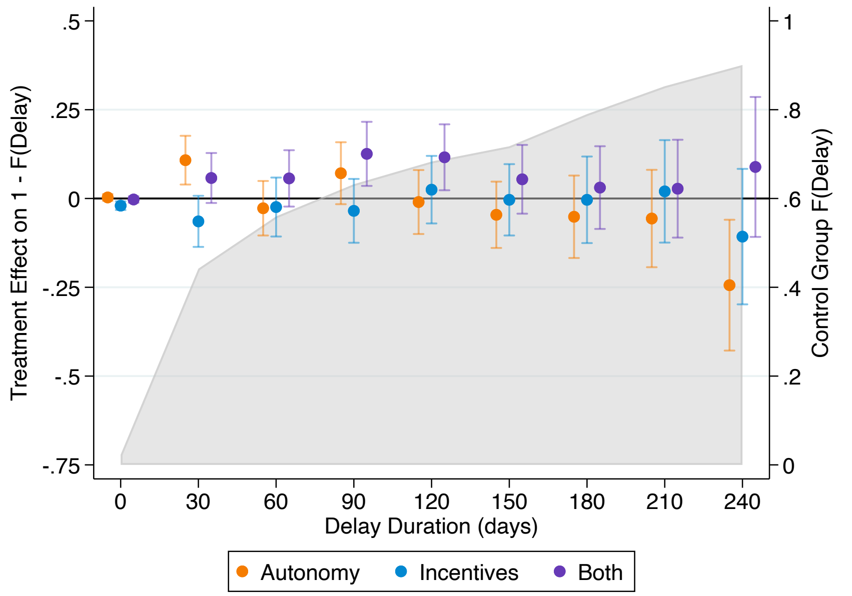
FIGURE 6: COST BENEFIT OF EXPERIMENT BY AG TYPE



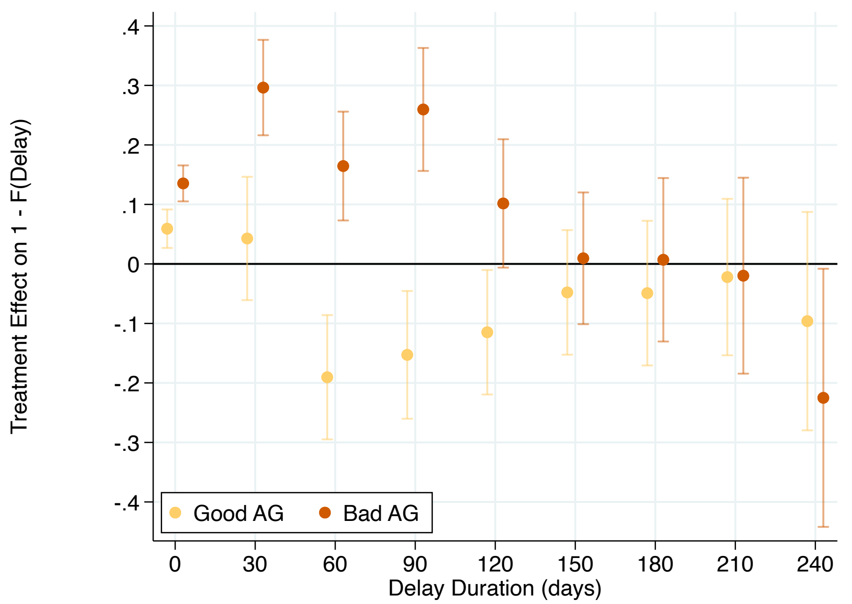
Notes: The figure shows the cost benefit of the experiments in districts with different levels of monitor alignment. The horizontal axis measures our proxy for the misalignment of a district's accountant general: the share of transactions approved in the last month of the fiscal year in the control group in year 1. Districts with a low AG June Share (low j_d) have more aligned monitors. The vertical axis measures the cumulative net savings by all districts with an accountant general who is less misaligned: $\sum_{d:j_d \leq x}\left[\left(\frac{-\eta_k(j_d)}{1+\eta_k(j_d)}\sum_{o\in d}\text{Expenditure}_{od}\times\text{Treatment}_o^k\right)-c_d\right]$ where $\eta_k(j_d)$ are estimated treatment effects of treatment k when monitor misalignment is j_d and c_d is the ex ante cost of performance pay bonuses to offices in district d (the number of offices in the district at each pay grade times the expected prize for each office). The figure shows large net savings for the incentives group at low levels of misalignment while net savings to the autonomy and both treatments only accrue at high levels of monitor misalignment.

FIGURE 7: TREATMENT EFFECTS ON APPROVAL DELAYS

Panel A: Overall Treatment Effects



Panel B: Autonomy Effect by AG Type



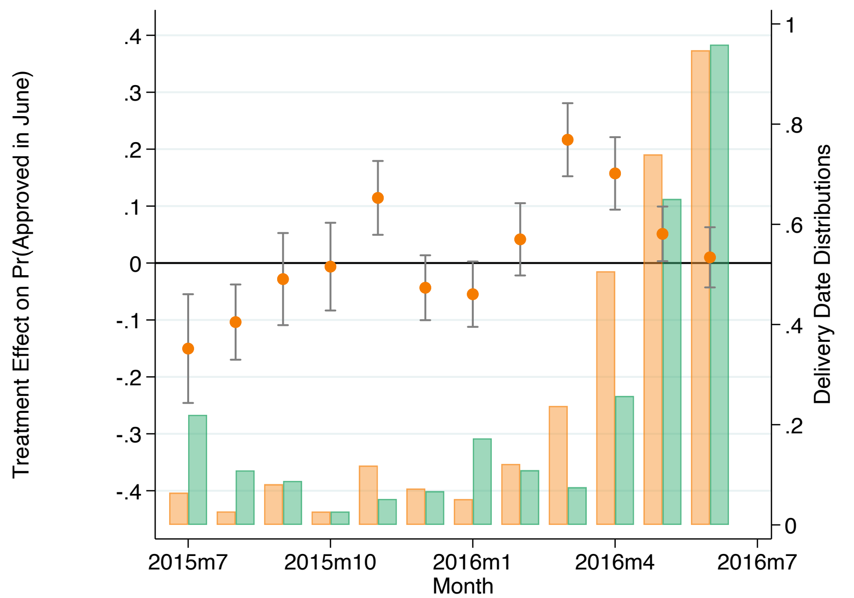
Notes: The figure shows the effects of the experiments on the delay between a purchased item's delivery and the approval of the purchase by the Accountant General (AG). Panel A shows a series of seemingly unrelated distributional regressions of the probability of delay of at least j days in year 2 normalized by the probability of a delay of at least j days in the control group in year 1 on treatment dummies, strata fixed effects γ_s and good fixed effects γ_g:

$$\frac{\mathbf{1}\left\{\text{delay}_{igo} \geq j\right\}}{\mathbb{P}\left(\text{delay} \geq j|\text{Control, Year1}\right)} = \alpha + \sum_{k=1}^{3} \eta_k \text{Treatment}_o^k + \gamma_s + \gamma_g + \varepsilon_{igo}$$

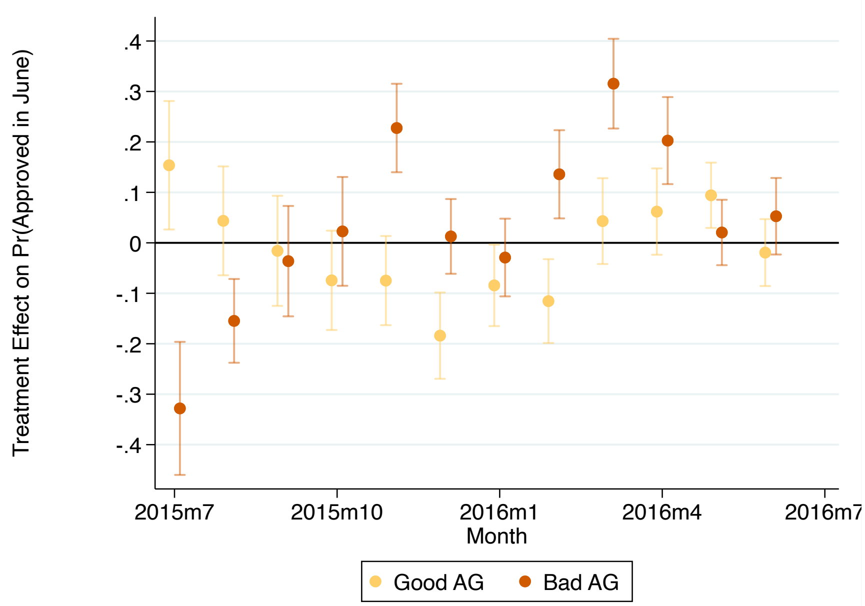
the panel also shows the CDF of delays in the control group in year 1 for reference. Panel B extends this regression to separately estimate treatment effects for good (June share of approvals ≤ 0.22) and bad (June share of approvals > 0.22) AGs.

FIGURE 8: EFFECTS OF AUTONOMY TREATMENT ON HOLD UP AT THE END OF THE FISCAL YEAR

Panel A: Overall Effect of Autonomy Treatment on Holdup



Panel B: Autonomy Effect by AG Type

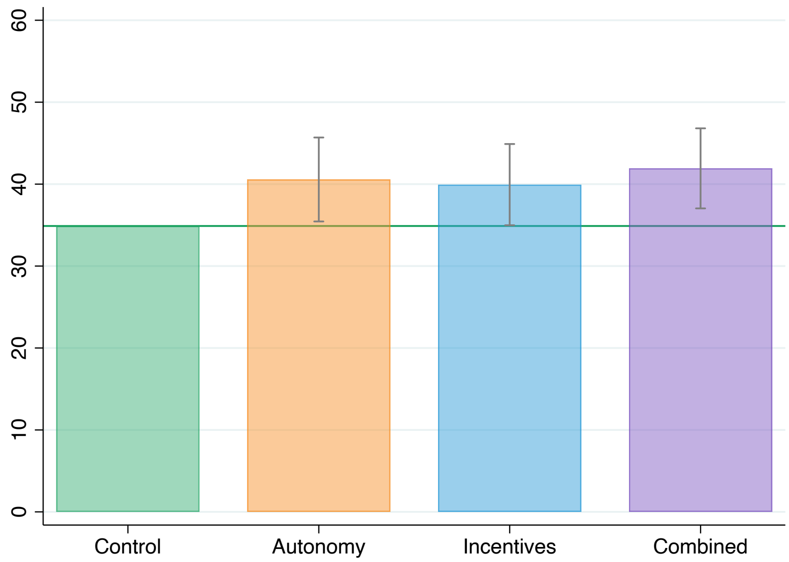


Notes: The figure shows the effects of the autonomy treatment on holdup by the AG at the end of the fiscal year. We focus on how the treatments change the probability that items purchased in different months are approved in June (the last month of the fiscal year).

$$\mathbf{1}\left\{\text{Approved in June}_{igo}\right\} = \alpha + \sum_{k=1}^{3}\sum_{m=Jul}^{Jun} \eta_{mk}\mathbf{1}\left\{\text{PurchaseMonth}_{igo} = m\right\} \times \text{Treatment}_{o}^{k} + \sum_{m=Jul}^{Jun} \gamma_m \mathbf{1}\left\{\text{PurchaseMonth}_{igo} = m\right\} + \gamma_g + \varepsilon_{igo}$$

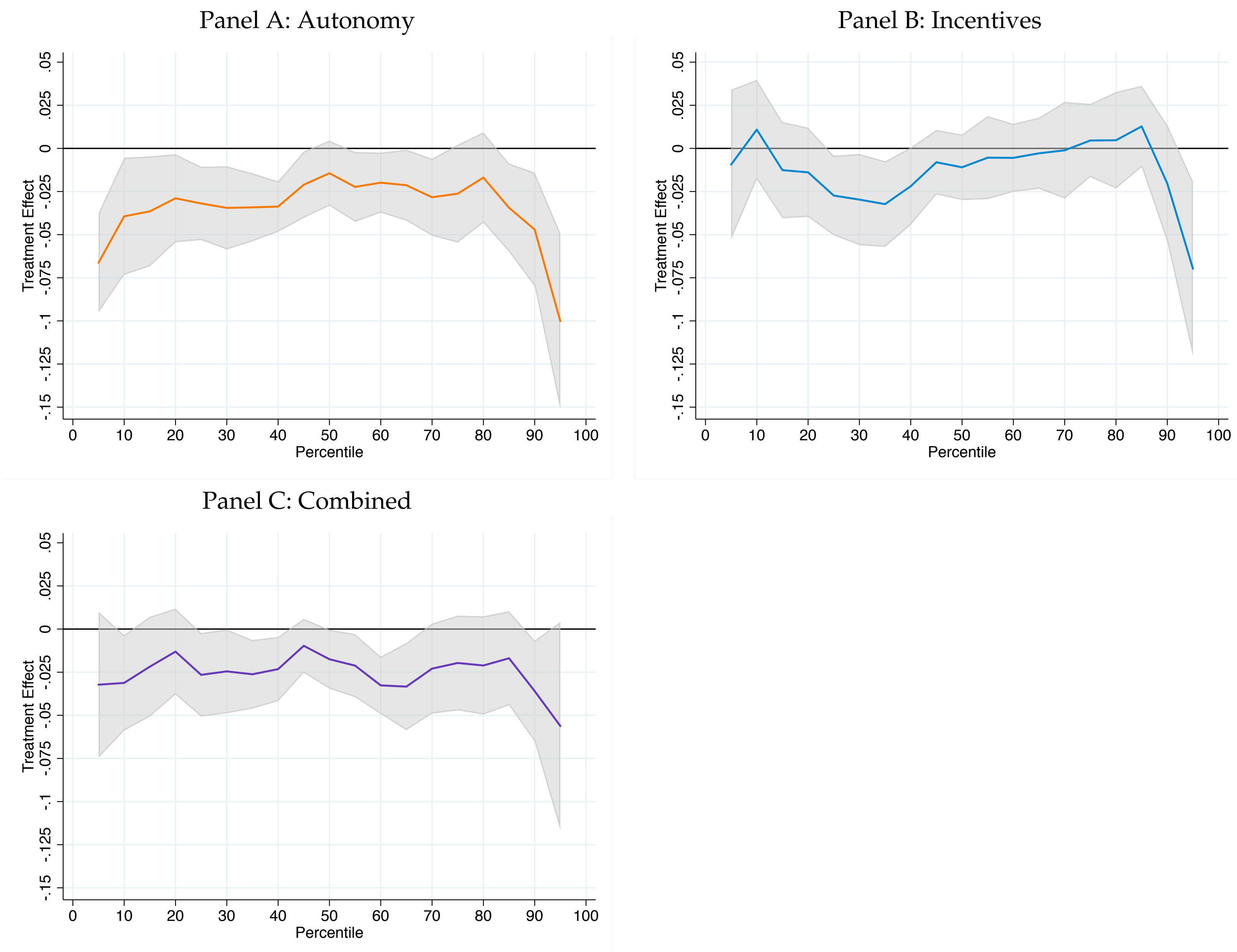
Panel A shows the η_{mk} coefficients for the autonomy treatment and also the raw distribution of delivery dates of purchases approved in June in the autonomy treatment (in orange) and control (in green) groups. Panel B runs the regression separately for less aligned (below median) and more aligned (above median) AGs.

FIGURE 9: TIME ALLOCATED TO PROCUREMENT



Notes: The figure shows our analysis of the effects of the experiment on the time bureaucrats allocate to procurement. Our endline survey asks bureaucrats to allocate months of the year to "very busy", "somewhat busy" and "not busy" months for procurement. The next question asks bureaucrats to specify the fraction of their time in each type of month they spend on procurement. We first combine these into a measure of the total amount of time in the year spent on procurement by averaging the answers to the latter question, weighting by the former. We find a 14% increase in the total amount of time spent on procurement in the incentives treatment, a 16% increase in the autonomy group, and a 20% increase in the combined group. We cannot reject the hypothesis that the increase is the same in all three groups ($p = 0.70$).

FIGURE 10: QUANTILE TREATMENT EFFECTS



Notes: The figure shows quantile treatment effects of the three treatments on prices paid. We use the specification used in table 2, controlling for the scalar measure of item variety. We estimate treatment effects from the 5th to the 95th percentile, in increments of 5.

TABLE 1: BALANCE ACROSS TREATMENT ARMS

	Control	Regression Coefficients			Joint Test
	mean/sd	Incentives	Autonomy	Both	All = 0
Office Characteristics					
	1.01	−0.007	0.033	0.012	2.360
Number of Public Bodies	{0.086}	(0.007)	(0.024)	(0.013)	[0.071]*
		[0.346]	[0.210]	[0.460]	[0.265]
	1.26	0.069	0.222	0.186	2.427
Number of Accounting Entities	{0.635}	(0.086)	(0.100)**	(0.087)**	[0.065]*
		[0.407]	[0.028]**	[0.038]**	[0.076]*
	0.39	−0.022	−0.009	−0.011	0.287
Share of June Approvals	{0.205}	(0.024)	(0.024)	(0.024)	[0.835]
		[0.363]	[0.693]	[0.649]	[0.828]
	1.13	0.003	0.027	0.077	1.089
# POs During Experiment	{0.403}	(0.047)	(0.051)	(0.050)	[0.353]
		[0.940]	[0.616]	[0.134]	[0.364]
District (χ^2 p-val)		[0.856]	[0.972]	[0.897]	[0.351]
Department (χ^2 p-val)		[0.168]	[0.958]	[0.858]	[0.639]
Procurement Officer Characteristics					
	52.03	−1.263	−0.493	0.345	1.109
Age	{6.883}	(0.938)	(0.952)	(0.875)	[0.345]
		[0.186]	[0.622]	[0.700]	[0.392]
	0.70	0.024	−0.011	0.007	0.137
Male	{0.460}	(0.056)	(0.058)	(0.056)	[0.938]
		[0.683]	[0.841]	[0.897]	[0.945]
	0.09	0.025	0.043	0.070	1.062
Bachelors Degree	{0.281}	(0.037)	(0.040)	(0.041)*	[0.365]
		[0.522]	[0.280]	[0.072]*	[0.354]
	0.76	0.004	−0.033	−0.013	0.179
Masters Degree	{0.429}	(0.054)	(0.056)	(0.055)	[0.910]
		[0.928]	[0.555]	[0.813]	[0.914]
	0.15	−0.029	−0.010	−0.058	0.791
Ph.D Degree	{0.362}	(0.044)	(0.046)	(0.042)	[0.499]
		[0.530]	[0.822]	[0.172]	[0.507]
	0.52	0.028	−0.083	−0.014	1.111
Pay Grade $\leq$ 18	{0.502}	(0.064)	(0.065)	(0.065)	[0.344]
		[0.660]	[0.192]	[0.822]	[0.320]

Continued on next page

Table 1 – *Continued from previous page*

	Control	Regression Coefficients			Joint Test
	mean/sd	**Incentives**	**Autonomy**	**Both**	**All = 0**
	0.33	−0.017	0.060	0.011	0.590
Pay Grade 19	{0.472}	(0.060)	(0.062)	(0.061)	[0.621]
		[0.790]	[0.324]	[0.852]	[0.617]
	0.15	−0.010	0.023	0.003	0.182
Pay Grade ≥ 20	{0.356}	(0.045)	(0.048)	(0.046)	[0.908]
		[0.804]	[0.612]	[0.940]	[0.908]
Year-1 Budget Shares					
	0.80	0.024	−0.004	0.009	0.594
Operating Expenses	{0.223}	(0.024)	(0.026)	(0.025)	[0.619]
		[0.328]	[0.875]	[0.708]	[0.611]
	0.03	−0.005	−0.004	−0.008	0.142
Physical Assets	{0.115}	(0.012)	(0.013)	(0.013)	[0.935]
		[0.664]	[0.769]	[0.546]	[0.944]
	0.05	0.005	−0.001	−0.003	0.394
Repairs & Maintenance	{0.098}	(0.010)	(0.010)	(0.010)	[0.757]
		[0.625]	[0.904]	[0.784]	[0.783]
	0.53	0.021	−0.001	−0.038	0.895
POPS Universe	{0.327}	(0.037)	(0.038)	(0.039)	[0.444]
		[0.579]	[0.971]	[0.352]	[0.467]
	0.15	0.027	0.025	−0.002	1.547
Analysis Sample	{0.173}	(0.020)	(0.021)	(0.018)	[0.201]
		[0.194]	[0.229]	[0.886]	[0.197]
Year-2 Budget Shares					
	0.78	−0.008	0.003	0.026	0.712
Operating Expenses	{0.240}	(0.027)	(0.028)	(0.027)	[0.545]
		[0.761]	[0.911]	[0.354]	[0.585]
	0.04	0.001	−0.019	−0.013	1.302
Physical Assets	{0.131}	(0.015)	(0.013)	(0.014)	[0.273]
		[0.971]	[0.140]	[0.368]	[0.302]
	0.05	0.001	0.000	−0.011	2.162
Repairs & Maintenance	{0.097}	(0.010)	(0.010)	(0.009)	[0.091]*
		[0.901]	[0.988]	[0.222]	[0.112]
	0.53	0.012	−0.001	−0.022	0.337
POPS Universe	{0.311}	(0.036)	(0.036)	(0.037)	[0.799]
		[0.716]	[0.973]	[0.529]	[0.790]

Continued on next page

Table 1 – *Continued from previous page*

	Control	Regression Coefficients			Joint Test
	mean/sd	**Incentives**	**Autonomy**	**Both**	**All = 0**
Analysis Sample	0.16 {0.196}	0.011 (0.023) [0.649]	0.007 (0.022) [0.725]	−0.018 (0.020) [0.388]	1.029 [0.379] [0.375]
Number of Offices	136	150	148	153	

Notes: The table shows balance of a range of covariates across the treatment arms. Each row of the table studies balance of a particular covariate. For continuous variables, the first column shows the mean and standard deviation (in curly brackets) of the variable in the control group. The next three columns show regression coefficients from a regression of the covariate on treatment indicators together with their robust standard errors in brackets, and the p-value from randomization inference on null hypothesis of no difference between that group and the control group. The final column shows the F-statistic on the joint test that no treatment group differs from the control group. Beneath it, we display its asymptotic p-value and beneath that its randomization inference p-value. To test whether the offices are equally distributed across departments and districts, we present p-values from Pearson's χ^2 tests for the equality of proportions.

TABLE 2: TREATMENT EFFECTS ON PRICES PAID AND GOOD VARIETY

	Variety			Unit Price				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Autonomy	0.016	0.010	-0.000	-0.085	-0.086	-0.080	-0.082	-0.085
	(0.030)	(0.023)	(0.009)	(0.038)	(0.032)	(0.031)	(0.034)	(0.038)
	[0.646]	[0.705]	[0.859]	[0.046]	[0.018]	[0.023]	[0.030]	[0.051]
Incentives	0.006	0.025	0.005	-0.016	-0.026	-0.022	-0.020	-0.017
	(0.030)	(0.023)	(0.009)	(0.038)	(0.030)	(0.033)	(0.034)	(0.038)
	[0.846]	[0.325]	[0.449]	[0.723]	[0.476]	[0.571]	[0.625]	[0.690]
Both	0.037	0.059	0.006	-0.070	-0.083	-0.072	-0.086	-0.070
	(0.030)	(0.023)	(0.009)	(0.041)	(0.032)	(0.033)	(0.039)	(0.041)
	[0.265]	[0.021]	[0.498]	[0.130]	[0.025]	[0.053]	[0.043]	[0.115]
Item Variety Measure	Scalar	Coarse	ML	None	Attribs	Scalar	Coarse	ML
p(All = 0)	0.660	0.080	0.809	0.168	0.054	0.093	0.087	0.178
p(Autonomy = Incentives)	0.749	0.537	0.494	0.146	0.077	0.119	0.119	0.138
p(Autonomy = Both)	0.461	0.031	0.573	0.741	0.927	0.807	0.932	0.761
p(Incentives = Both)	0.302	0.144	0.957	0.262	0.133	0.227	0.136	0.238
Observations	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771

Notes: The table shows the overall treatment effects of the three treatments. The table shows estimates of equation (2):

$$y_{igto} = \alpha + \sum_{k=1}^{3} \eta_k \text{Treatment}_o^k + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$$

where y_{igto} is the outcome of interest in purchase i of good g at time t by office o. In columns 1–3 it is the scalar (column 1), coarse (column 2) or machine learning (column 3) measure of good variety, while in columns 4–8 it is the log unit price. Treatment_o^k indicates the three treatment groups; q_{igto} is the quantity purchased to capture good-specific bulk discounts; δ_s and γ_g are stratum and good fixed effects, respectively; and $\mathbf{X}_{igto}$ are purchase-specific controls. In column 5 we control for the full set of good attributes, while the remaining columns control for the scalar (column 6), coarse (column 7), or machine learning (column 8) measures of good variety. We weight regressions by expenditure shares in the control group so that treatment effects can be interpreted as effects on expenditure, and the residual term ε_{igto} is clustered at the cost center level. Below each coefficient we report standard errors clustered by cost center in parentheses, and p-values from randomization inference tests of the hypothesis that the treatment has no effect on any office in square brackets.

TABLE 3: TIME USE AND PROCUREMENT PERFORMANCE

	First Stage		Quantification		Placebo	
	(1) Bad AG	(2) Good AG	(3) Autonomy	(4) Incentives	(5) Good AG	(6) Bad AG
Time Spent on Procurement			-0.012	-0.010		
			(0.006)	(0.005)		
			[0.001]	[0.001]		
Autonomy	6.975				3.785	
	(2.867)				(4.930)	
	[0.037]				[0.510]	
Incentives		8.335				-0.330
		(2.464)				(4.338)
		[0.004]				[0.950]
First-stage F statistic	5.92	11.44			0.59	0.01
Observations	6,273	6,355	6,273	6,355	3,454	2,201

Notes: The table quantifies how much of the estimated effect comes from additional time spent working on procurement on offices' procurement performance. The first column estimates the effect of the autonomy treatment on time spent on procurement when the Accountant General (AG) is relatively misaligned, for whom figure 5 reveals a significant effect of the treatment on prices. In this sample, we estimate equation (2):

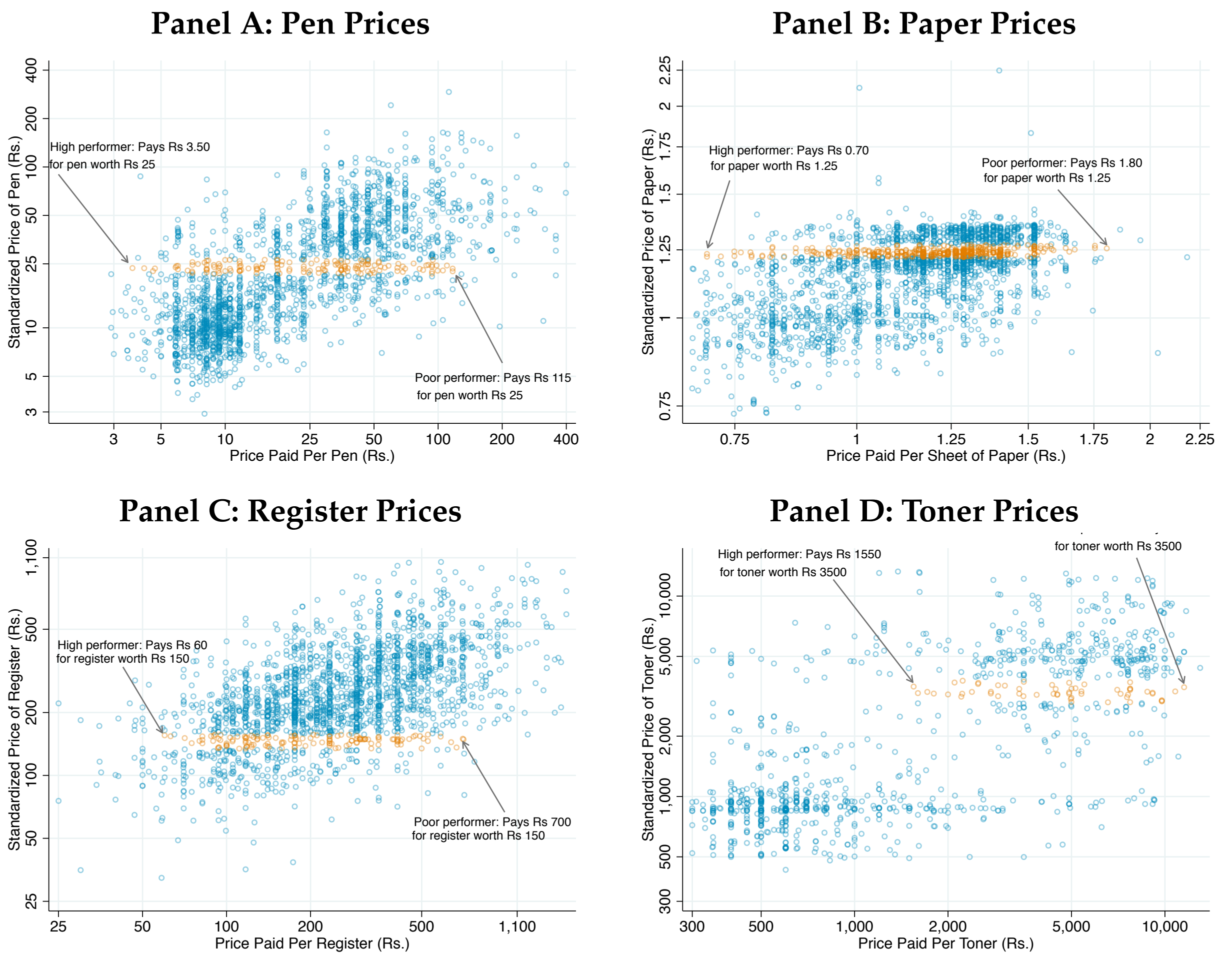
$$y_{igto} = \alpha + \sum_{k=1}^{3} \eta_k \text{Treatment}_o^k + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$$

where y_{igto} is the time spent on procurement and terms are defined as above. Analogously, column 2 estimates the effect of the incentive treatment on time spent on procurement when the AG is relatively well aligned. Columns 3 and 4 use this as a first stage to quantify the extent to which the experiment's effects on prices occur through changes in time use. Note that these estimates capture the price impacts of all actions the PO takes that are correlated with the time spent on procurement, they should not be interpreted as causal effects of time on prices. Columns 5 and 6 explore, as a placebo, the effect of the experimental treatments on time use in the regions where figure 5 shows no effect of the experiment on prices. Correspondingly, there is no effect on time use. Standard errors clustered by office are in parentheses. p-values from randomization inference under the null hypothesis of no effect for any office are in square brackets.

Web Appendix (Not For Publication)

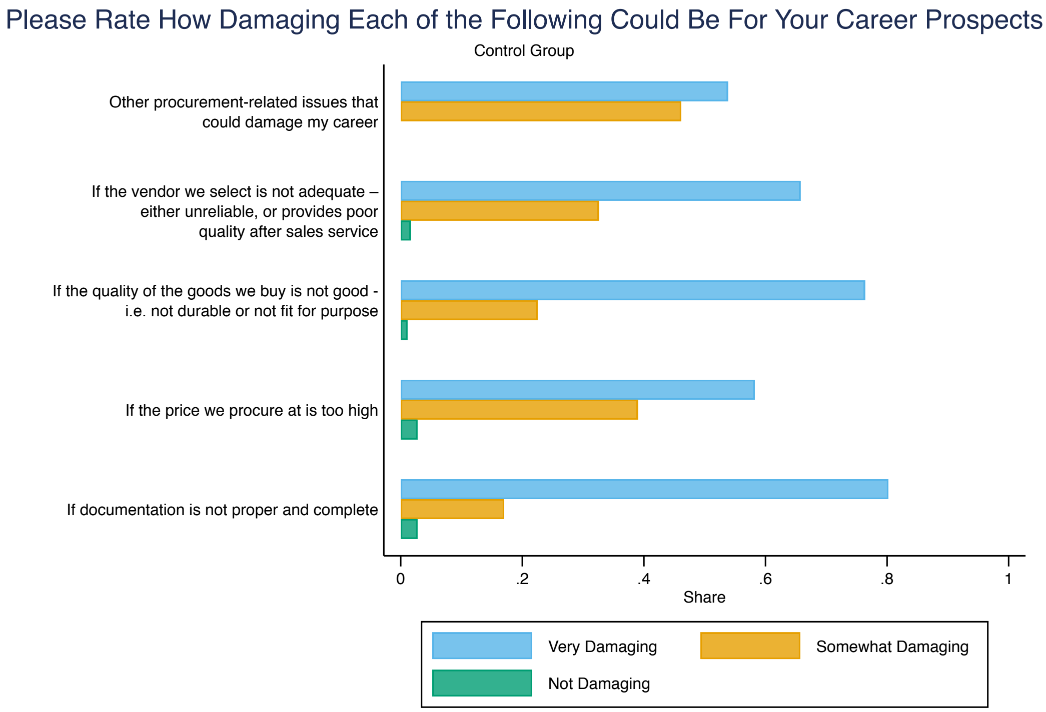
A Supplementary Figures and Tables

Figure A.1: Prices Paid Vary Wildly. Even for the Same Variety of Item



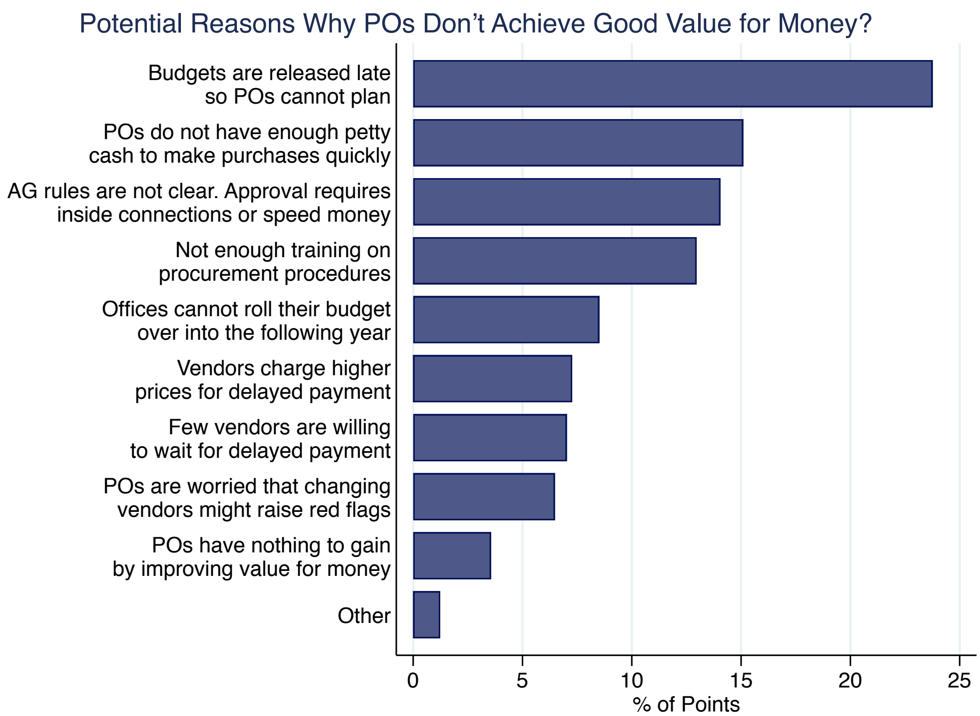
Notes: The figure shows the distribution of unit prices and standardized prices for four of the homogeneous items in our data. Each circle in the figures is a purchase. The horizontal axes display the actual price paid, while the vertical axes display the standardized prices using the scalar item variety measure described in section 5.1. Intuitively, this measure is our prediction of how much the item would have cost on average if it had been purchased in the control group, a standardized measure of the item's variety. The orange circles highlight a set of purchases with the same standardized value, illustrating the striking heterogeneity in prices even for the same item.

FIGURE A.2: HOW POOR PROCUREMENT PERFORMANCE CAN DAMAGE CAREERS



Notes: The figure shows responses among the control group in the endline survey to a question asking them about whether various types of poor performance in procurement could damage their careers. Each bar shows the share of respondents picking that option.

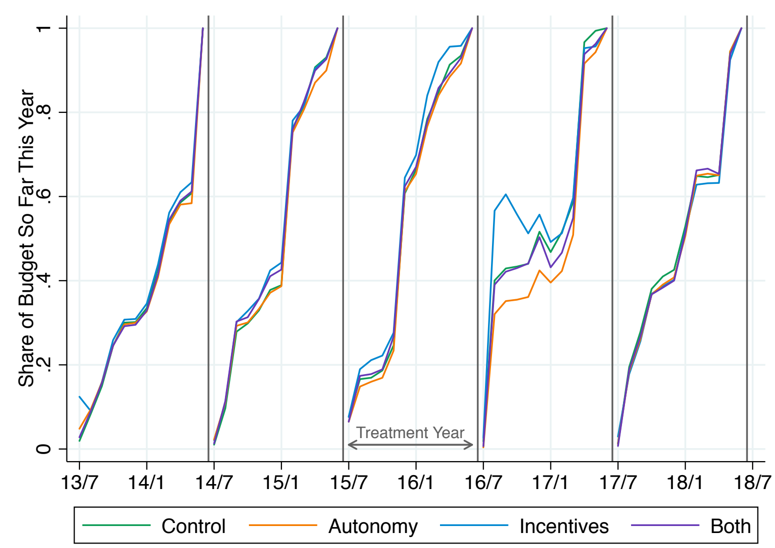
FIGURE A.3: CONTROL GROUP REASONS FOR LOW VALUE FOR MONEY



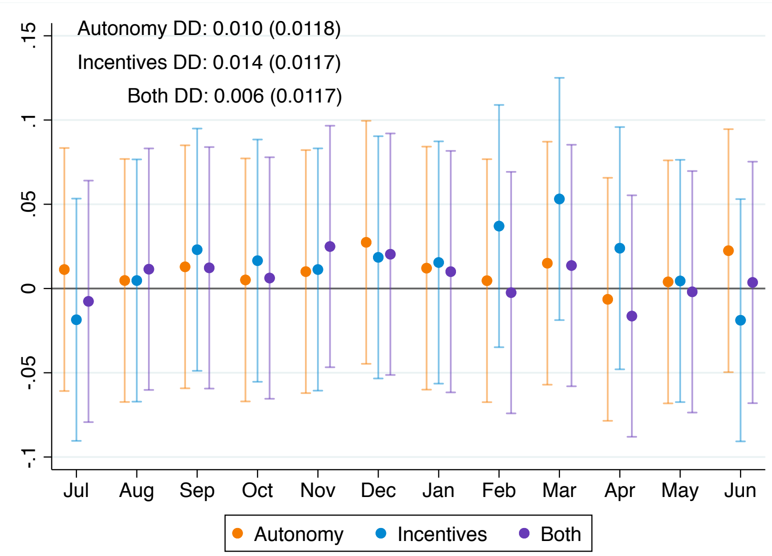
Notes: The figure shows responses among the control group in the endline survey to a question asking them about the reasons they felt that value for money was not being achieved in public procurement. Respondents were asked to allocate 100 points among the 10 options in proportion to how important they thought each option was. Each bar shows the mean number of points allocated to that option.

Figure A.4: Budget Release Timing Unaffected

Panel A: Share of Budget Released Over Time



Panel B: Difference in Differences Estimates

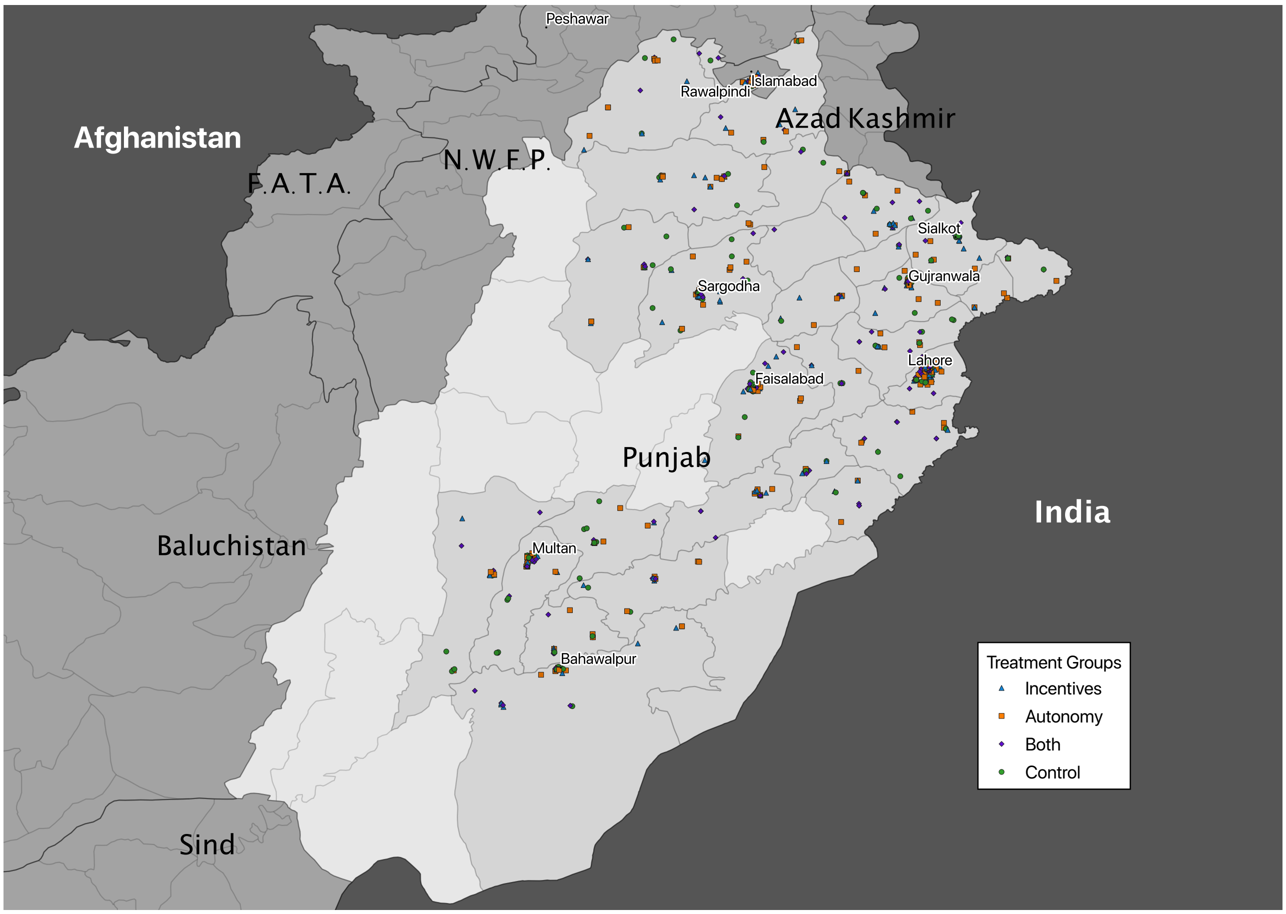


Notes: The figure shows that the timing of budget releases to the offices in the study was unaffected. A third component of the autonomy treatment attempted to improve the frequency and regularity of budget releases, but it was not possible to implement this. Panel A shows how the average share of offices' annual budget evolves over each year in each treatment group. The treatment year (July 2015–June 2016) does not look visibly different from the other years, and any slight differences from other years appear to have affected all four groups in the same way. Panel B shows estimates of the η_{km} coefficients from a differences in differences estimation of

$$s_{ot} = \sum_{k=1}^{3} \sum_{m=Jul}^{Jun} \eta_{km} \text{Treatment}_o^k \times \mathbf{1}\left\{\text{Month of year} = m\right\} \times \mathbf{1}\left\{\text{Fiscal Year 2015–16}\right\} + \delta_t + \gamma_o + \varepsilon_{ot}$$

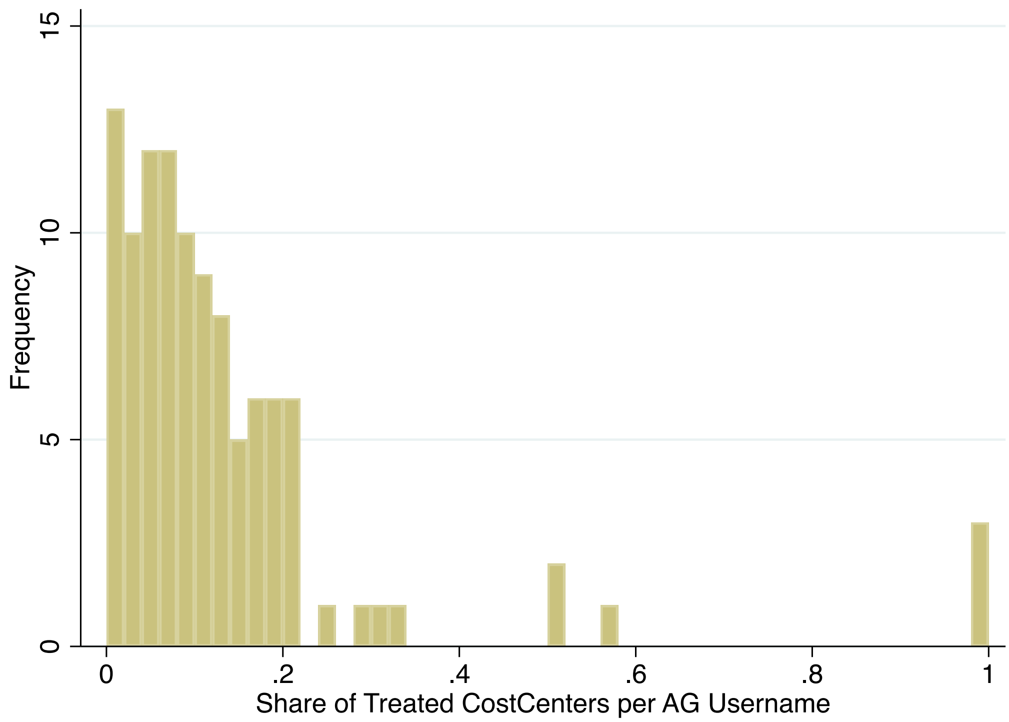
where s_{ot} is the share of office o's annual budget that has been released to it by month t, δ_t are month fixed effects, γ_o are office fixed effects and ε_{ot} are residuals. Overlaid on the figure are estimates of difference in difference coefficients of the average effect in the 2015–16 fiscal year in each treatment group.

FIGURE A.5: LOCATION OF SAMPLE OFFICES



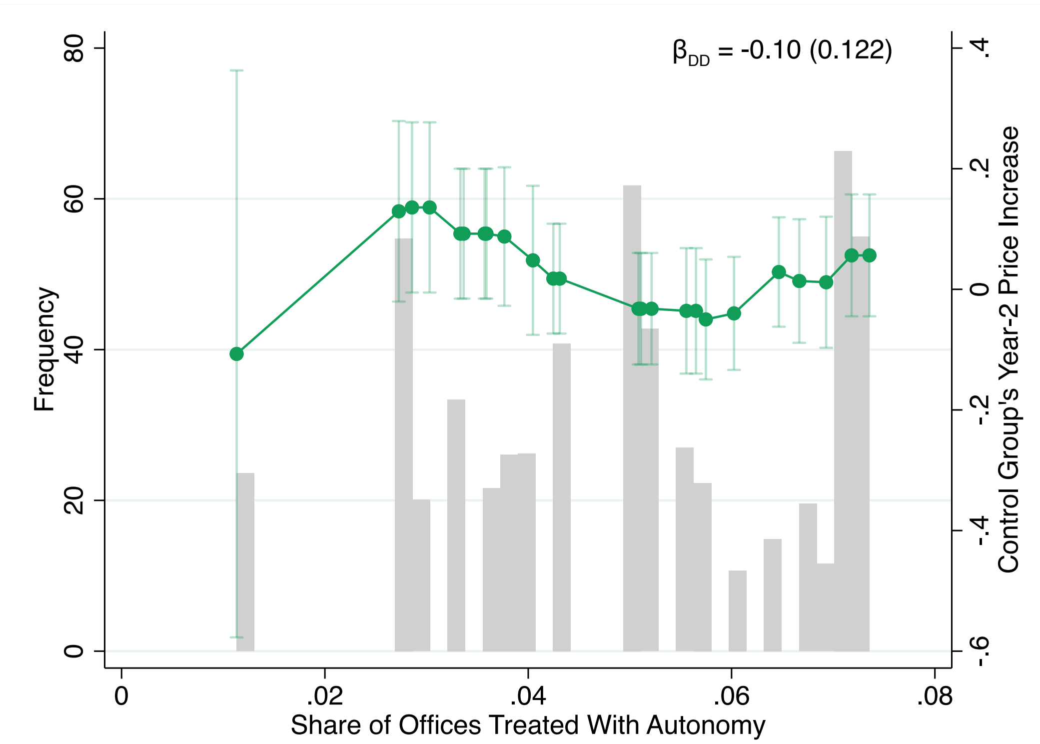
Notes: The figure shows the location of the offices in the study. The offices are located in 26 of the 36 districts in Punjab. Green dots denote control offices, orange dots the autonomy group, blue dots the performance pay group, and purple dots the combined treatment.

FIGURE A.6: SAMPLE OFFICES ARE A SMALL SHARE OF THE OFFICES OVERSEEN BY USERS AT THE ACCOUNTANT GENERAL'S OFFICE



*Notes:*Each transaction approved by the accountant general's office is associated with a particular officer's username. The figure shows the share of cost centers associated with each username that are in the treated groups of our experiment. The figure shows that for the vast majority of users at the accountant general's office, fewer than 20% of their offices are treated.

FIGURE A.7: PRICE CHANGES IN THE CONTROL GROUP ARE NOT LARGER WHEN MORE OFFICES RECEIVE THE AUTONOMY TREATMENT

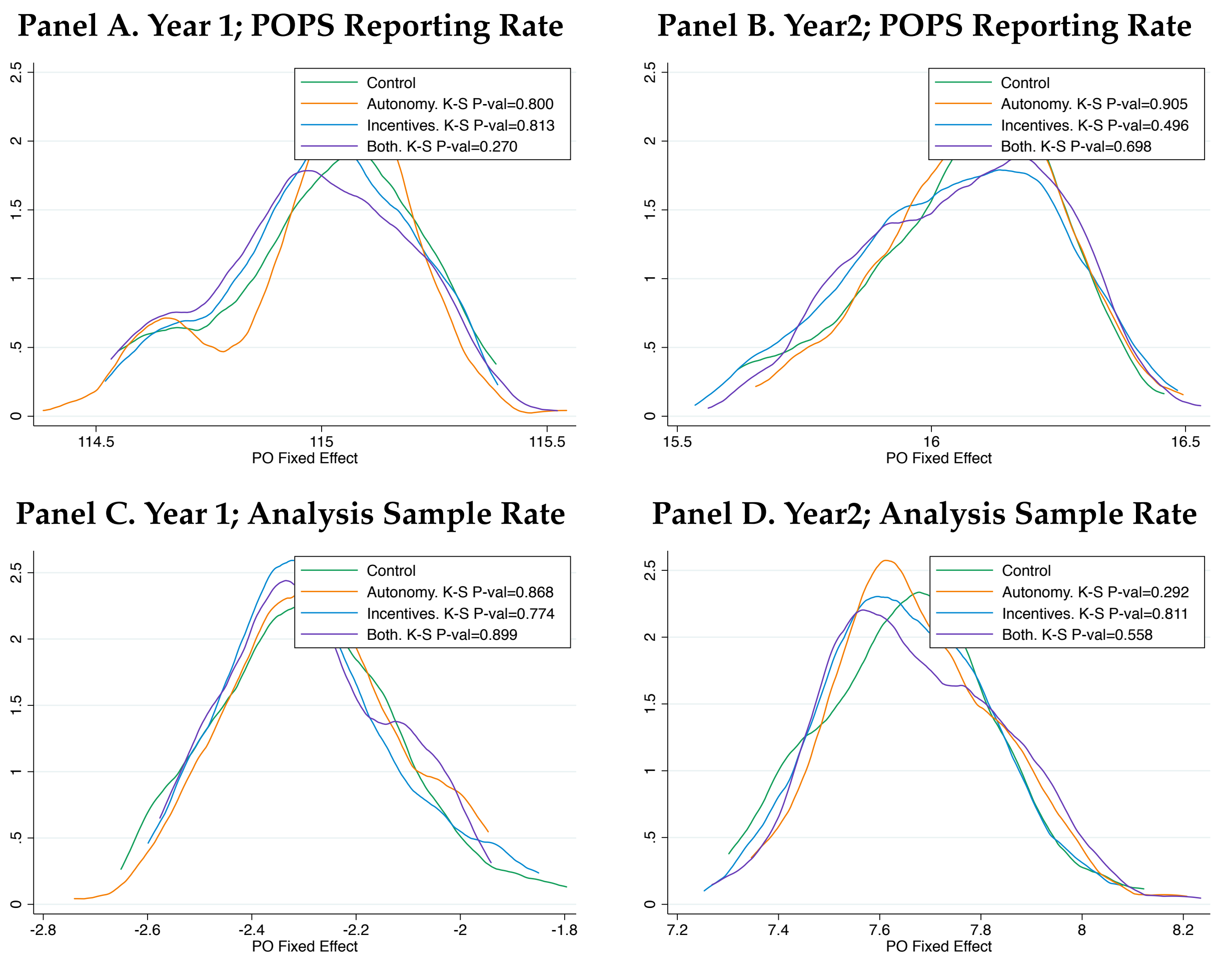


Notes: The figure shows how prices change between year 1 (before the rollout of the autonomy treatment) and year 2 (after the rollout) in offices in the control group as a function of the share of the offices monitored by an accountant general that receive the autonomy treatment. For each accountant general's office, we run the regression $p_{igto} = \alpha\hat{v}^{\text{scalar}}_{igto} + \beta_{Y2}\text{Year2}_t + \gamma_g + \rho_g q_{igto} + \varepsilon_{igto}$, where $\hat{v}_{igto}$ is the scalar measure of item variety, in a sample of control group procurement offices supervised by an accountant general with a share of offices in the autonomy group within 0.01 of the office in question. The figure presents these estimates with their 95% confidence intervals in green. We also overlay on the picture the difference in differences estimate of β_{DD} in the following regression

$$p_{igto} = \alpha\hat{v}^{\text{scalar}}_{igto} + \beta_{Y2}\text{Year2}_t + \beta_{DD}\text{Year2}_t \times \text{AutonomyShare}_o + \gamma_g + \rho_g q_{igto} + \delta_g t + \varepsilon_{igto}$$

where AutonomyShare_o is the share of procurement officers monitored by the same accountant general as officer o who receive the autonomy treatment and the regression is run only amongst procurement officers in the control group.

FIGURE A.8: BALANCE OF THE DISTRIBUTION OF ATTRITION RATES ACROSS OFFICES



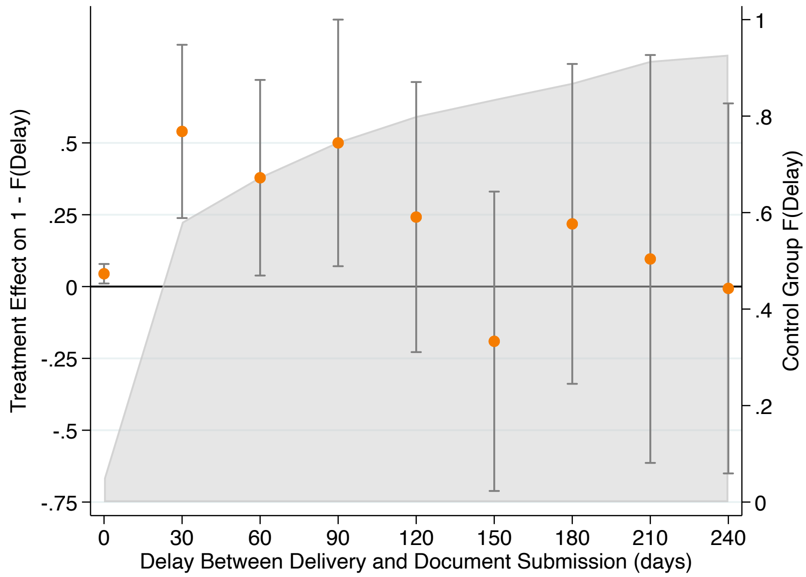
Notes: The figure shows the distribution of procurement office fixed effects δ_o in regressions of the form

$$s_{bco} = \mathbf{X}_{bco}\beta + \gamma_c + \delta_o + \varepsilon_{bco}$$

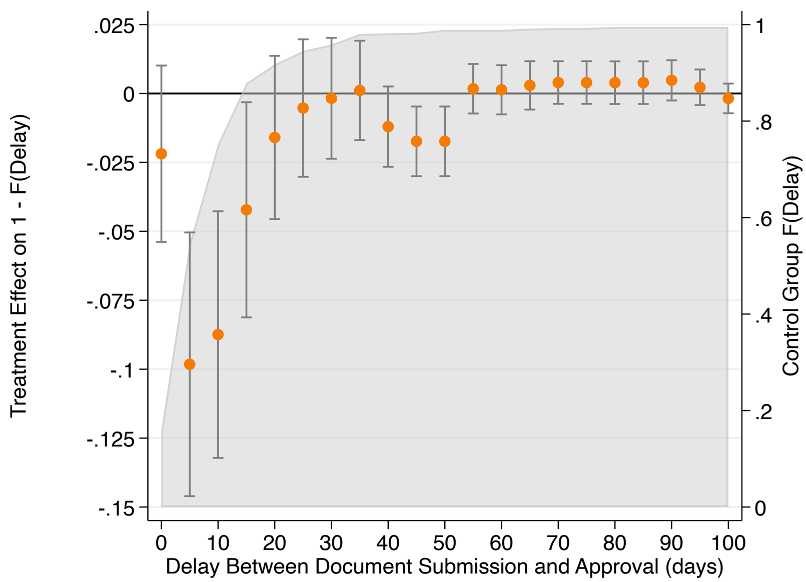
where s_{bco} is the share of a transaction (bill) b by office c in an accounting code o that is reported in POPS (panels A and B) or that is represented in our analysis sample (panels C and D); $\mathbf{X}_{bco}$ are quadratic time and bill amount controls, γ_c are accounting code fixed effect, δ_o are procurement office fixed effects, and ε_{bco} is an error term. Panels A and C use bills from year 1 of the experiment, while panels B and D analyze year 2. The panels show kernel density estimates of the distributions of the procurement office fixed effects in the 3 treatment groups and the control group. The panels also show exact P-values form Kolmogorov-Smirnov tests of the equality of each treatment group's distribution and the control group's.

FIGURE A.9: DECOMPOSING AUTONOMY EFFECTS ON APPROVAL DELAYS

Panel A: Delay Between Delivery and Document Submission



Panel B: Delay Between Document Submission and Approval



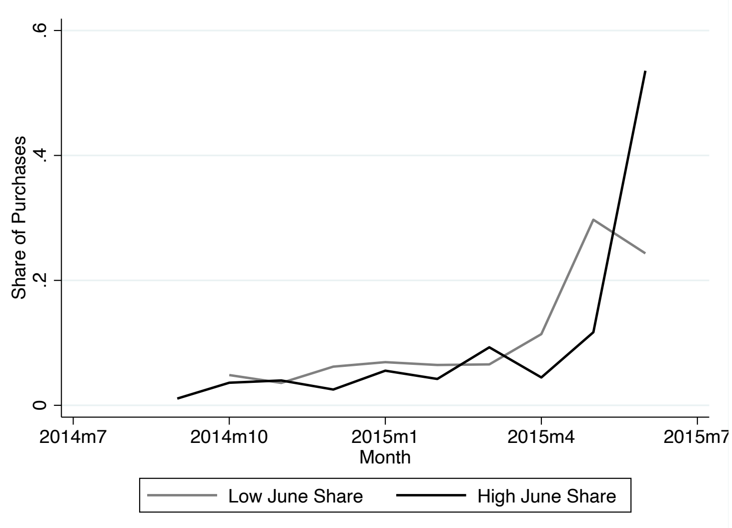
Notes: The figure decomposes the effects of the autonomy treatment on the delay between a purchased item's delivery and the approval of the purchase by the Accountant General (AG) into the delay between the item's delivery and the submission of the documents for approval (Panel A) and the delay between the document's submission and their approval by the AG (Panel B). The estimates come from a series of seemingly unrelated distributional regressions of the probability of delay of at least j days in year 2 normalized by the probability of a delay of at least j days in the control group in year 1 on treatment dummies, strata fixed effects γ_s and good fixed effects γ_g:

$$\frac{\mathbf{1}\left\{\text{delay}_{igo} \geq j\right\}}{\mathbb{P}\left(\text{delay} \geq j|\text{Control, Year1}\right)} = \alpha + \sum_{k=1}^{3} \eta_k \text{Treatment}_o^k + \gamma_s + \gamma_g + \varepsilon_{igo}$$

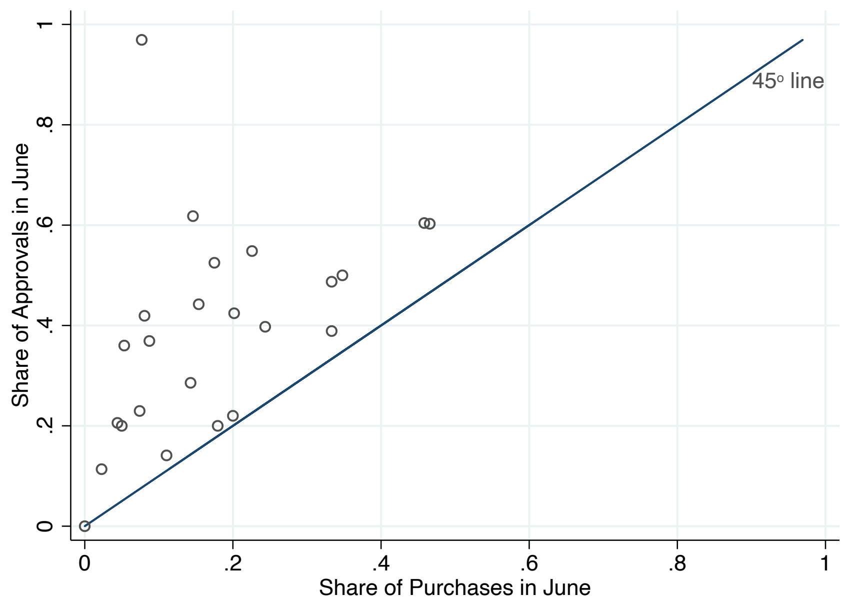
the panel also shows the CDF of delays in the control group in year 1 for reference.

FIGURE A.10: VARIATION IN JUNE APPROVAL RATES

Panel A: High and Low Approval Rate Districts



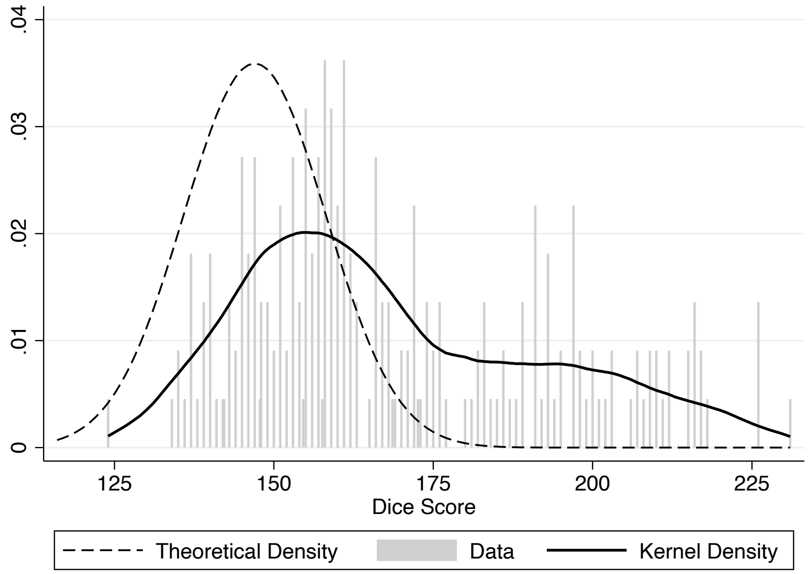
Panel B: Sources of Variation in June Approval Rates



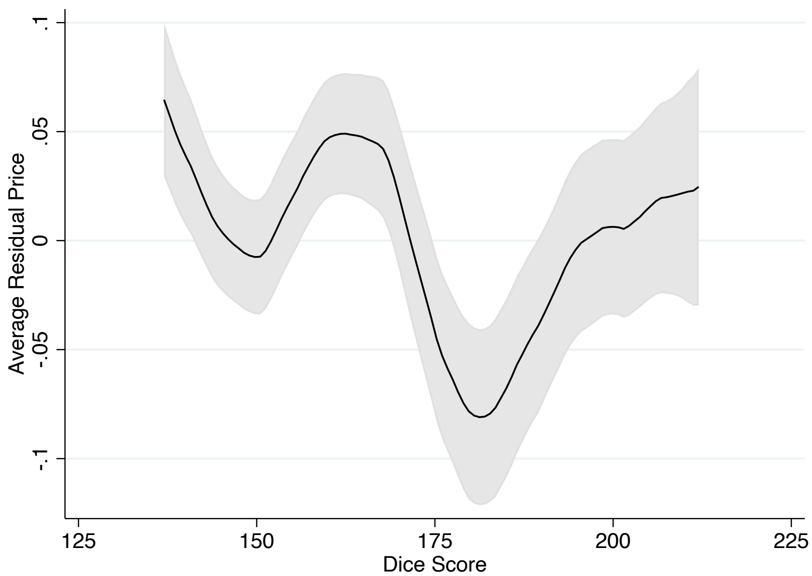
Notes: The figure shows the variation in our proxy for AG type, the share of approvals done in June. Panel A compares the approval rates in districts with high (above median) and low (below median) shares of transactions approved in June. Panel B shows the variation across districts' AG offices in the share of transactions made in June (the last month of the fiscal year) and the share of transactions approved in June (our proxy for the misalignment of the AG). Both aggregates are calculated in the control group in year 1.

Figure A.11: Dice Scores as a Proxy for PO Type do not Predict Year 1 Prices

Panel A: Distribution of Dice Scores



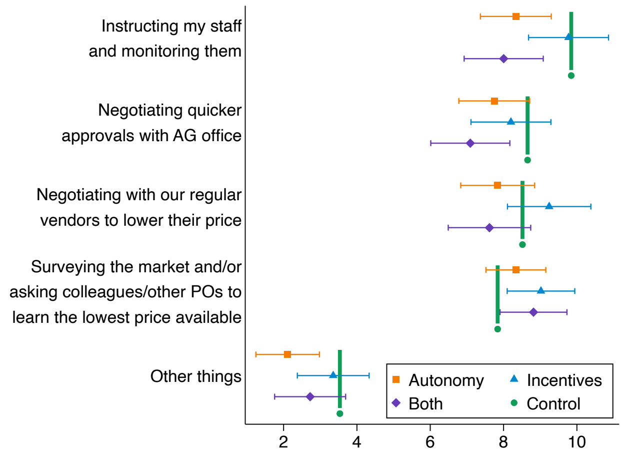
Panel B: Dice Scores do not Predict Year 1 Prices



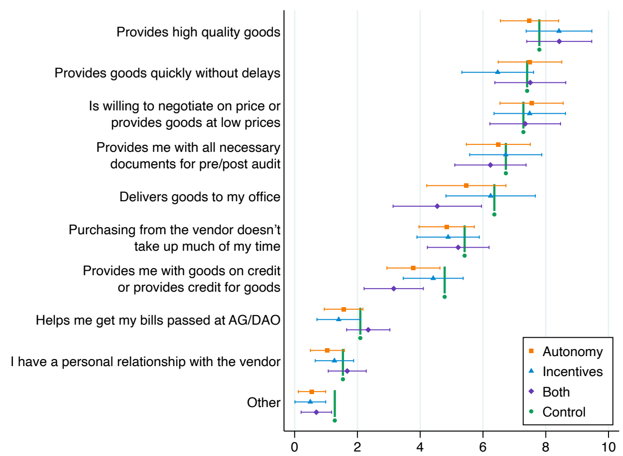
Notes: The figure shows that the dice scores in the lab in the field measure of dishonesty studied in Fischbacher & Föllmi-Heusi (2013) and Hanna & Wang (2017) are a poor proxy for PO type in our setting. The dice scores come from a game in which subjects privately roll a die 42 times and report each roll. In each roll they are free to report the number either on the top or the bottom of the die. Subjects play against each other and those achieving the highest scores win prizes. The dashed line in panel A shows the theoretical distribution of the total scores if a fair die is rolled 42 times. The histogram and the solid line (kernel density) show the totals achieved by our subjects. Panel B shows a semi-parametric regression of log unit prices in year 1 in the control group and the autonomy group on controls and the dice scores, showing that the dice scores do not predict prices in year 1. Together, the findings in panel A and B suggest that while there is significant variation in the dice scores in our sample, it is not predictive of procurement performance and hence is a poor proxy for PO type in our setting.

FIGURE A.12: TIME ALLOCATION ACROSS PROCUREMENT TASKS

Panel A: Reducing Amount Paid for Goods



Panel B: Choosing Vendors



Notes: The figure shows analysis of the responses to our endline survey questions on mechanisms. The panels show differences (and their 95% confidence intervals) in mean responses across the 4 treatment arms, weighting offices by the number of purchases they make. The control group mean is in green, autonomy in orange, incentives in blue, and combined in purple. Panel A shows responses to the question *"Of all the time you spend trying to reduce the total amount your cost center pays (including hidden costs) for the goods you want, what percentage of your time do you and your staff spend on each of the tasks below?"* Panel B shows responses to *"Please think about the vendors you currently make contingent purchases from, and the vendors you could potentially make contingent purchases from. Which of the following characteristics of vendors are important to you in deciding which vendor(s) to buy from?"* The possible responses are shortened to fit in the figures. The full text of the responses is in the questionnaire in the Social Science Registry at https://www.socialscienceregistry.org/trials/610. Answers may not sum to 100 since respondents seem in many cases to have interpreted the questions to mean percentage of total time rather than percentage of time spent *on procurement*.

Table A.1: Universe of Generic Goods Accounting Codes

Code	Category	Description
Panel A: A03 Operating Expenses		
A03004	Other	Furnace Oil - Non Operational
A03070		Others
A03170	Fees	Others
A03204	Communication	Electronic Communication
A03205		Courier And Pilot Service
A03206		Photography Charges
A03270		Others
A03304	Utilities	Hot And Cold Weather
A03305		POL For Generator
A03370		Others
A03401	Occupancy Costs	Charges
A03405		Rent Other Than Building
A03408		Rent Of Machine & Equipment
A03410		Security
A03470		Others
A03501	Operating Leases	Machinery And Equipment
A03502		Buildings
A03503		Motor Vehicles
A03504		Computers
A03506		Medical Machinery And Technical Equipment
A03570		Others
A03901	General	Stationery
A03902		Printing And Publication
A03904		Hire Of Vehicles
A03905		Newspapers Periodicals And Books
A03907		Advertising & Publicity
A03919		Payments To Others For Service Rendered
A03921		Unforeseen Exp. For Disaster Preparedness

Continued on next page

Table A.1 – *Continued from previous page*

Code	Category	Description
A03927		Purchase Of Drug And Medicines
A03933		Service Charges
A03940		Unforeseen Expenditure
A03942		Cost Of Other Stores
A03955		Computer Stationary
A03970		Others
A03971		Cost Of State Trading Medicines
A03972		Expenditure On Diet For Patient
A03978		Free Text Books
	Panel B: A09 Physical Assets	
A09105	Purchase of Physical Assets	Transport
A09107		Furniture And Fixtures
A09108		Livestock
A09170		Others
A09204	Computer Accessories	License Fee For Software
A09302	Commodity Purchases	Fertilizer
A09303		Coal
A09370		Others
A09401	Other Stores and Stock	Medical Stores
A09402		Newsprint
A09403		Tractors
A09404		Medical And Laboratory Equipment
A09405		Workshop Equipment
A09406		Storage And Carrying Receptacles
A09407		Specific Consumables
A09408		Generic Consumables
A09409		Medical Stocks
A09410		Life Saving Medical Supplies
A09411		General Utility Chemicals
A09412		Specific Utility Chemicals
A09413		Drapery Fabrics Clothing And Allied Materials

Continued on next page

Table A.1 – *Continued from previous page*

Code	Category	Description
A09414		Insecticides
A09470		Others
A09501		Transport
A09502	Transport	Diplomatic Cars
A09503		Others
A09601		Plant And Machinery
A09602	Plant & Machinery	Cold Storage Equipment
A09603		Signalling System
A09604		Railways Rolling Stock
A09701	Furniture & Fixtures	Furniture And Fixtures
A09702		Unkempt Furnishings
A09801		Livestock
A09802	Livestock	Purchase Of Other Assets - Others
A09803		Meters & Services Cables
A09899		Others
	Panel C: A13 Repairs and Maintenance	
A13101	Machinery & Equipment	Machinery And Equipment
A13199		Others
A13201	Furniture & Fixture	Furniture And Fixture
A13370	Buildings & Structure	Others
A13470	Irrigation	Others
A13570	Embankment & Drainage	Others
A13701		Hardware
A13702	Computer Equipment	Software
A13703		I.T. Equipment
A13920	Telecommunication	Others

TABLE A.2: PROJECT TIMELINE

Year 1: July 2014 – June 2015	
06/14	Cost Centers allocated to treatment arms
07–08/14	Trainings on POPS and treatment brochures
08–09/14	Follow-up trainings on POPS
02/15	Performance Evaluation Committee midline meeting
05–06/15	AG checklist rolled out
Year 2: July 2015 – June 2016	
07–10/15	Refresher trainings on treatments and POPS
10/15	Higher cash balance rolled out
04/16	Performance Evaluation Committee midline meeting
06/16	Experiment ends
Post-Experiment	
08-09/16	Endline survey part 1 & Missing data collection
02/17	Performance Evaluation Committee endline meeting
02–03/17	Endline survey part 2

TABLE A.3: BALANCE OF ATTRITION OF ITEMS

	All Generics				Analysis Objects			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Incentives	0.006	0.003	-0.003	0.005	0.009	0.005	-0.002	0.006
	(0.015)	(0.017)	(0.013)	(0.012)	(0.018)	(0.020)	(0.015)	(0.015)
Autonomy	-0.011	-0.009	-0.009	-0.003	-0.010	0.000	-0.008	-0.001
	(0.016)	(0.016)	(0.013)	(0.012)	(0.018)	(0.019)	(0.015)	(0.015)
Both	-0.038*	-0.013	-0.017	-0.001	-0.041*	-0.013	-0.020	-0.002
	(0.018)	(0.018)	(0.014)	(0.013)	(0.020)	(0.020)	(0.016)	(0.017)
Assets: Fertilizer	0.000	0.000	0.000	0.000				
	(.)	(.)	(.)	(.)				
Assets: General Utility Chemicals	-0.061	-0.108*	0.019	-0.014				
	(0.053)	(0.053)	(0.022)	(0.019)				
Assets: Insecticides	0.111	-0.174***	-0.019**	-0.011				
	(0.067)	(0.049)	(0.007)	(0.006)				
Assets: Lab Equipment	-0.263***	-0.422***	0.069**	0.066*				
	(0.055)	(0.046)	(0.026)	(0.029)				
Assets: Other Commodity	0.073	-0.053	-0.019	-0.020*	0.000	0.000	0.000	0.000
	(0.093)	(0.068)	(0.012)	(0.009)	(.)	(.)	(.)	(.)
Assets: Other Stocks and Stores	-0.068	-0.188	0.044	0.009				
	(0.138)	(0.150)	(0.036)	(0.015)				
Assets: Purchase of Furniture & Fixture	-0.108	-0.248***	0.047*	0.104***	-0.167	-0.132	0.081***	0.168***
	(0.067)	(0.066)	(0.019)	(0.021)	(0.114)	(0.097)	(0.020)	(0.031)
Assets: Purchase of Plant & Machinery	-0.273***	-0.420***	0.079***	0.039	-0.301**	-0.341***	0.122***	0.078**

	(0.071)	(0.079)	(0.021)	(0.025)	(0.111)	(0.094)	(0.022)	(0.027)
Assets: Purchase of Transport	-0.288***	-0.442***	0.032	0.087***				
	(0.061)	(0.051)	(0.029)	(0.020)				
Assets: Specific Utility Chemicals	-0.055	-0.282***	0.008	0.037**	-0.120	-0.199*	0.031	0.077**
	(0.084)	(0.073)	(0.010)	(0.012)	(0.123)	(0.092)	(0.017)	(0.024)
OpEx: Advertising	-0.124*	-0.314***	0.217***	0.238***	-0.203	-0.266***	0.232***	0.254***
	(0.058)	(0.046)	(0.023)	(0.023)	(0.105)	(0.073)	(0.026)	(0.025)
OpEx: Courier	-0.455***	-0.735***	-0.055	-0.139**				
	(0.090)	(0.062)	(0.049)	(0.042)				
OpEx: Electricity	0.138*	-0.135**	0.495***	0.437***	0.055	-0.090	0.506***	0.450***
	(0.061)	(0.046)	(0.027)	(0.025)	(0.105)	(0.073)	(0.027)	(0.025)
OpEx: Elextronic Communication	-0.382***	-0.678***	-0.000	-0.088*				
	(0.092)	(0.101)	(0.037)	(0.039)				
OpEx: Medicines	-0.196***	-0.422***	0.134***	0.119***				
	(0.055)	(0.045)	(0.014)	(0.015)				
OpEx: Newspapers	0.147*	-0.156***	0.289***	0.309***	0.070	-0.107	0.301***	0.324***
	(0.064)	(0.046)	(0.022)	(0.024)	(0.107)	(0.073)	(0.022)	(0.024)
OpEx: Other	0.009	-0.256***	0.197***	0.177***	-0.065	-0.209**	0.214***	0.194***
	(0.055)	(0.043)	(0.015)	(0.016)	(0.105)	(0.072)	(0.018)	(0.018)
OpEx: Other Stores	-0.148**	-0.366***	0.070***	0.058***	-0.212*	-0.310***	0.093***	0.080***
	(0.055)	(0.043)	(0.015)	(0.013)	(0.104)	(0.072)	(0.016)	(0.015)
OpEx: Other Stores: Computer/Stationery	0.090	-0.167**	0.367***	0.371***	0.014	-0.118	0.385***	0.388***
	(0.070)	(0.061)	(0.050)	(0.048)	(0.112)	(0.084)	(0.049)	(0.047)
OpEx: Other Utilities	-0.245***	-0.420***	0.071*	0.137	-0.339**	0.123	0.066**	0.590***
	(0.058)	(0.103)	(0.033)	(0.082)	(0.104)	(0.110)	(0.025)	(0.133)

OpEx: Payments for Services	-0.298***	-0.574***	0.058***	-0.009				
	(0.054)	(0.043)	(0.015)	(0.015)				
OpEx: Printing	-0.044	-0.270***	0.173***	0.125***	-0.120	-0.219**	0.190***	0.143***
	(0.054)	(0.045)	(0.016)	(0.019)	(0.104)	(0.073)	(0.019)	(0.020)
OpEx: Rent not on Building	-0.437***	-0.604***	0.003	0.020				
	(0.064)	(0.069)	(0.021)	(0.024)				
OpEx: Rent of Machine	-0.443***	-0.625***	-0.007	0.023				
	(0.065)	(0.069)	(0.021)	(0.023)				
OpEx: Stationery	0.076	-0.138**	0.352***	0.372***	0.002	-0.091	0.369***	0.389***
	(0.056)	(0.042)	(0.018)	(0.015)	(0.104)	(0.072)	(0.019)	(0.020)
Repairs: Computer Hardware	-0.155*	-0.304***	0.107**	0.116**	-0.237	-0.249*	0.124**	0.136**
	(0.079)	(0.086)	(0.041)	(0.045)	(0.121)	(0.100)	(0.041)	(0.045)
Repairs: Computer Software	-0.328***	-0.538***	0.042	-0.019				
	(0.058)	(0.088)	(0.021)	(0.017)				
Repairs: Furniture & Fixtures	-0.380***	-0.651***	-0.006	-0.077***	-0.459***	-0.606***	0.009	-0.063***
	(0.055)	(0.043)	(0.015)	(0.015)	(0.103)	(0.072)	(0.015)	(0.016)
Repairs: IT Equipment	-0.220	-0.053	0.085	0.199***	-0.290	0.018	0.103	0.230***
	(0.123)	(0.167)	(0.066)	(0.040)	(0.153)	(0.170)	(0.068)	(0.040)
Repairs: Machinery & Equipment	-0.321***	-0.569***	0.020	-0.026	-0.399***	-0.521***	0.035*	-0.009
	(0.055)	(0.044)	(0.016)	(0.015)	(0.104)	(0.072)	(0.016)	(0.016)
Repairs: Other Building	-0.142**	-0.485***	0.150***	0.058*				
	(0.053)	(0.052)	(0.012)	(0.026)				
Date	-0.007	-0.001***	0.004	-0.000***	-0.005	-0.001***	0.006	-0.000***
	(0.006)	(0.000)	(0.006)	(0.000)	(0.007)	(0.000)	(0.007)	(0.000)
Date^2	0.000	0.000***	-0.000	0.000***	0.000	0.000***	-0.000	0.000***

	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
log Amount	-0.121***	-0.132***	-0.108***	-0.144***	-0.082**	-0.095**	-0.101***	-0.128***
	(0.028)	(0.020)	(0.023)	(0.024)	(0.027)	(0.032)	(0.025)	(0.031)
log(Amount)^2	0.004***	0.005***	0.004***	0.005***	0.002	0.002	0.004**	0.004**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
Assets: Generic Consumables		-0.400***		0.131***				
		(0.051)		(0.019)				
Constant	69.447	13.868***	-41.798	6.610***	47.408	15.965***	-60.546	7.598***
	(61.980)	(1.333)	(63.492)	(0.944)	(69.733)	(1.531)	(66.118)	(1.075)
Observations	23,423	22,498	23,423	22,498	17,361	16,553	17,361	16,553
R^2	0.33	0.33	0.28	0.32	0.25	0.24	0.24	0.27
Year	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
Reporting Share	POPS	POPS	Analysis	Analysis	POPS	POPS	Analysis	Analysis

Table A.4: Heterogeneity of Treatment Effects by Procurement Officer Dice Score

	(1)	(2)	(3)	(4)
Autonomy	0.2791	0.4386	0.3442	0.4123
	(0.2820)	(0.2396)	(0.2317)	(0.2589)
	[0.396]	[0.134]	[0.213]	[0.180]
Incentives	-0.0413	0.2079	0.0963	0.1967
	(0.3089)	(0.2457)	(0.2574)	(0.2774)
	[0.915]	[0.505]	[0.770]	[0.579]
Both	-0.0431	0.2665	0.1409	0.1225
	(0.4106)	(0.3199)	(0.3319)	(0.3965)
	[0.915]	[0.504]	[0.717]	[0.797]
Autonomy $\times$ Dice Score	-0.0023	-0.0033	-0.0026	-0.0030
	(0.0017)	(0.0015)	(0.0014)	(0.0016)
	[0.249]	[0.071]	[0.122]	[0.112]
Incentives $\times$ Dice Score	0.0001	-0.0015	-0.0007	-0.0013
	(0.0019)	(0.0015)	(0.0016)	(0.0017)
	[0.954]	[0.426]	[0.698]	[0.541]
Both $\times$ Dice Score	-0.0003	-0.0022	-0.0013	-0.0013
	(0.0025)	(0.0019)	(0.0020)	(0.0024)
	[0.918]	[0.336]	[0.579]	[0.648]
Item Variety Control	None	Attribs	Scalar	Coarse
p(All Interactions = 0)	0.167	0.056	0.156	0.132
Observations	10,283	10,283	10,283	10,283

Notes: The table shows heterogeneity of treatment effects by the degree of misalignment of the procurement officer, as measured by their score in the dice game measure of dishonesty studied in Fischbacher & Föllmi-Heusi (2013) and Hanna & Wang (2017) and summarized in appendix figure A.11. We estimate treatment effect heterogeneity by interacting our proxy for PO type $\hat{\mu}_o$ with treatment dummies $p_{igto} = \alpha + \eta \text{Autonomy}_o + \zeta \text{Autonomy}_o \times \hat{\mu}_o + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$.

B Proofs and Other Theory Material

Proof of Proposition 1

By re-arranging Δ_A, we can see that $\Delta_A > 0$ if and only if

$$\frac{\theta_{AG}}{1-\theta_{AG}} < \frac{\frac{\theta_{PO}}{1-\theta_{PO}}\left(p_M - p_{MA}\right)}{-\left(\frac{\theta_{PO}}{1-\theta_{PO}}\left(p_M - p_{MM}\right) + \left(c - p_{AM}\right)\right)}$$

Both the numerator and the denominator are positive. Hence for any θ_{PO} there exists $\bar{\theta}_{AG}\left(\theta_{PO}\right) > 0$ such that for all $\theta_{AG} > \bar{\theta}_{AG}\left(\theta_{PO}\right)$, the inequality holds. Note that $\bar{\theta}_{AG}\left(\theta_{PO}\right)$ is increasing in θ_{PO}.

If the bandit competition effect is absent ($p_M = p_{MM}$) and the bad-monitor effect is low with respect to the good-monitor effect ($(p_{AM} - c)/(p_M - p_{MA}) < 1$), then the inequality becomes

$$\frac{\theta_{AG}}{1-\theta_{AG}} > \frac{\theta_{PO}}{1-\theta_{PO}}\frac{p_M - p_{MA}}{p_{AM} - c} > \frac{\theta_{PO}}{1-\theta_{PO}},$$

and there exists $\bar{\theta}_{AG}\left(\theta_{PO}\right) \in (0,1)$ such that it holds as an equality.

Proof of Corollary 1

Autonomy treatment decreases price if

$$\frac{\theta_{AG}}{1-\theta_{AG}} > \frac{\frac{\theta_{PO}}{1-\theta_{PO}}\left(p_M - p_{MA}\right)}{-\left(\frac{\theta_{PO}}{1-\theta_{PO}}\left(p_M - p_{MM}\right) + \left(c - p_{AM}\right)\right)},$$

or

$$\frac{\theta_{AG}}{1-\theta_{AG}}\left(-\left(\frac{\theta_{PO}}{1-\theta_{PO}}\left(p_M - p_{MM}\right) + \left(c - p_{AM}\right)\right)\right) > \frac{\theta_{PO}}{1-\theta_{PO}}\left(p_M - p_{MA}\right).$$

If $\theta_{AG} = \theta_{PO}$, the condition becomes:

$$-\left(\frac{\theta_{PO}}{1-\theta_{PO}}\left(p_M - p_{MM}\right) + \left(c - p_{AM}\right)\right) > p_M - p_{MA}.$$

As $p_M - p_{MM} < 0$, the condition is satisfied a fortiori if

$$p_{AM} - c > p_M - p_{MA.}$$

which can be re-written as

$$p_M - p_{AM} < p_{MA} - c,$$

which is always true because

$$p_M - p_{AM} < p_{MM} - p_{AM} \leq p_{MA} - c,$$

where the first inequality is because $p_M < p_{MM}$ and the second inequality is true because of Complementarity Between Agent Types.

Proof of Proposition 2

The effect of the incentive treatment on price is:

$$\begin{aligned}\Delta_I &= \theta_{AG} p_{AM} + (1-\theta_{AG})\, c \\ &\quad - \left(\theta_{PO}\theta_{AG} p_{MM} + \theta_{PO}\left(1-\theta_{AG}\right) p_{MA} + \left(1-\theta_{PO}\right)\theta_{AG} p_{AM} + \left(1-\theta_{PO}\right)\left(1-\theta_{AG}\right) c\right) \\ &= \theta_{PO}\theta_{AG}\left(p_{AM} - p_{MM}\right) + \theta_{PO}\left(1-\theta_{AG}\right)\left(c - p_{MA}\right) < 0\end{aligned}$$

We also see that

$$\lim_{\theta_{AG}\to 1, p_{MM}\to p_{AM}} \Delta_I = 0$$

Statement and Proof of Proposition 3

Proposition 3. *(i) The price reduction generated by the combined autonomy and treatment effect is at least as large as the larger price reduction generated by Autonomy and Incentive as individual treatments.*

(ii) There exist values of $(\theta_{PO}, \theta_{AG})$ for which the weak inequality in (i) holds as a strict inequality.

Proof. The combined effect is

$$\begin{aligned}\Delta_C &= c - \left(\theta_{PO}\theta_{AG} p_{MM} + \theta_{PO}\left(1-\theta_{AG}\right) p_{MA} + \left(1-\theta_{PO}\right)\theta_{AG} p_{AM} + \left(1-\theta_{PO}\right)\left(1-\theta_{AG}\right) c\right) \\ &= \theta_{PO}\theta_{AG}\left(c - p_{MM}\right) + \theta_{PO}\left(1-\theta_{AG}\right)\left(c - p_{MA}\right) + \left(1-\theta_{PO}\right)\theta_{AG}\left(c - p_{AM}\right)\end{aligned}$$

Compare with

$$\begin{aligned}\Delta_I &= \theta_{AG} p_{AM} + (1-\theta_{AG})\, c \\ &\quad - \left(\theta_{PO}\theta_{AG} p_{MM} + \theta_{PO}\left(1-\theta_{AG}\right) p_{MA} + \left(1-\theta_{PO}\right)\theta_{AG} p_{AM} + \left(1-\theta_{PO}\right)\left(1-\theta_{AG}\right) c\right) \\ &= \theta_{PO}\theta_{AG}\left(p_{AM} - p_{MM}\right) + \theta_{PO}\left(1-\theta_{AG}\right)\left(c - p_{MA}\right) < 0\end{aligned}$$

and

$$\Delta_A = \theta_{PO}\theta_{AG}\left(p_M - p_{MM}\right) + \theta_{PO}\left(1-\theta_{AG}\right)\left(p_M - p_{MA}\right) + \left(1-\theta_{PO}\right)\theta_{AG}\left(c - p_{AM}\right)$$

The comparison with the incentive treatment yields:

$$\begin{aligned}\Delta_C - \Delta_I &= \theta_{PO}\theta_{AG}\left(c - p_{MM}\right) + \left(1-\theta_{PO}\right)\theta_{AG}\left(c - p_{AM}\right) - \theta_{PO}\theta_{AG}\left(p_{AM} - p_{MM}\right)\\ &= \theta_{PO}\theta_{AG}\left(c - p_{AM}\right) + \left(1-\theta_{PO}\right)\theta_{AG}\left(c - p_{AM}\right) < 0\end{aligned}$$

The comparison with the autonomy treatment yields:

$$\begin{aligned}\Delta_C - \Delta_A &= \theta_{PO}\theta_{AG}\left(c - p_{MM}\right) + \theta_{PO}\left(1-\theta_{AG}\right)\left(c - p_{MA}\right)\\ &\quad - \left(\theta_{PO}\theta_{AG}\left(p_M - p_{MM}\right) + \theta_{PO}\left(1-\theta_{AG}\right)\left(p_M - p_{MA}\right)\right)\\ &= \theta_{PO}\theta_{AG}\left(c - p_M\right) + \theta_{PO}\left(1-\theta_{AG}\right)\left(c - p_M\right) < 0\end{aligned}$$

For (ii), simply notice that for $\left(\theta_{PO}, \theta_{AG}\right) \in \{(0,0), (1,0), (0,1)\}$, either $\Delta_C = \Delta_I$, $\Delta_C = \Delta_A$, or both. $\square$

C Alternative, Micro-founded Model

This appendix presents an alternative model of the setting we study and the effects we expect from the experimental treatments. The model is still stylized, but instead of the primitives of the model being prices, the model's primitives are the procurement officer and the monitor's utility functions. In addition, both the procurement officer and the monitor have a continuum of types, giving rise to a continuum of prices. Nevertheless, the model remains a parsimonious framework that delivers highly stylized predictions to guide the analysis.

C.1 Set-up

This simple model describes our context, where procurement decisions are taken by an *officer* and monitored by a *monitor* with veto power.

For each purchase, the officer selects a *mark-up* $x \geq 0$. The mark-up x captures different forms of misalignment between the interests of the officer and her principal, the taxpayer. It can be interpreted as active waste (bribes), passive waste (inefficiency), or a combination of both. We will discuss both interpretations below.

The officer operates under a monitoring agency. The purchase is audited by the monitor with probability $1 - a$ (where a stands for autonomy – the probability that the officer is not audited). The purchase price is thus

$$p = c + x + \omega (1 - a),$$

where c is the cost of the good, x is the officer's mark-up, and ω is an additional cost introduced by the monitor.

If a purchase is audited, the officer receives a punishment proportional to the markup x. Finally, the officer faces an incentive to spend less. Her utility is:

$$u = \gamma \ln x - \mu (1 - a) x - bx,$$

where: the first term is the benefit the officer receives from the mark-up, which is scaled by γ, the weight the officer puts on her private utility; the second term is the cost the officer incurs if she is audited on the procured good, which depends on the effectiveness of the monitoring process, μ; and b in the third term represents the strength of a monetary incentive scheme whereby the officer is rewarded for spending less.

The model has two interpretations. In the *active* waste interpretation, the officer receives a bribe from the supplier in exchange for increasing the purchase price above the supply cost. The underlying assumption is that there is a bribing technology that transforms a mark-up x into a benefit for the officer $\gamma \ln x$. In this interpretation a higher markup has three effects: it increases the price of the purchased good by x; it produces utility for the officer, who enjoys the bribe, given by $\gamma \ln x$, and it imposes a risk of sanction on the officer given by $\mu (1-a) x$.

In the *passive* waste interpretation, the officer is lazy and prefers not to exert effort to locate the cheapest supplier or wring the lowest price from the chosen supplier. The underlying assumption is that there is a search/bargaining technology that transforms a mark-up x into a benefit for the officer $\gamma \ln x$: less work leads to higher prices. In this interpretation a higher mark-up has three effects too: it increases the price of the purchased good by x; it produces utility for the officer, who enjoys the lower effort, given by $\gamma \ln x$, and it imposes a risk of sanction on the officer given by $\mu (1-a) x$. Of course, it is also possible to interpret the model as a mix of active and passive waste.

The role of the monitor can also be interpreted in two ways. In the active waste interpretation, the monitor also receives a bribe and that raises the purchase price by $\omega (1-a)$. The monitor also punishes the officer for accepting bribes through $\mu (1-a) x$. In the passive waste interpretation, the monitor too dislikes effort: if there is an audit he may add to the price of good by taking a long time to process the purchase (perhaps because suppliers predict that it will take them a long time to be paid). This too raises the purchase price by $\omega (1-a)$. The monitor also punishes the officer for engaging in passive waste through $\mu (1-a) x$.

In both interpretations the monitor has a positive effect and a negative effect. The positive effect consists in disciplining the officer through $\mu (1-a) x$. As we shall see shortly, this induces the officer to decrease her mark-up x. The negative effect instead operates through $\omega (1-a)$: it is the additional passive or active waste that the monitor generates. The rest of the analysis will show that the overall effect of the monitor will depend on relative size of these two effects.

We now proceed with the analysis (normalizing c to zero without loss of generality). The officer selects the optimal mark-up level given her preference parameters and the environment she faces:

$$x = \frac{\gamma}{\mu (1-a) + b}$$

and the price is

$$p = \frac{\gamma}{\mu (1-a) + b} + \omega (1-a)$$

The price formula embodies the autonomy tradeoff: the first term captures the monitor's disciplining effect on the officer, while the second represents the additional mark-up imposed by the monitor.

This simple model thus captures the trade-off at the heart of the allocation of authority: giving more autonomy to the officer (higher a) increases markups especially if the officer puts a large weight on her private benefits γ, but it reduces supervision costs at the same time.

C.2 Treatment effects

Our two experimental treatments involve an increase in autonomy (higher a) and an increase in the power of incentives (higher b). The effects of the two treatments on prices (in percentage terms) are as follows

Proposition 4. *(i) An increase in autonomy decreases p if and only if ω is sufficiently large relative to γ, and the decrease is larger when ω is large*

(ii) An increase in incentive power always decreases p, but the decrease is larger when ω is small and tends to zero as $\omega \to \infty$.

Proof. For (i):

$$\frac{\frac{\partial p}{\partial a}}{p} = \frac{\frac{\partial}{\partial a}\left(\frac{\gamma}{\mu(1-a)+b} + (1-a)\,\omega\right)}{p} = \frac{\frac{\gamma\mu}{(\mu(1-a)+b)^2} - \omega}{p} < 0 \text{ iff } \omega > \bar{\omega} \equiv \frac{\gamma\mu}{\left(\mu\,(1-a)+b\right)^2}$$

Clearly $\frac{\frac{\partial p}{\partial a}}{p}$ is decreasing in ω and $\lim_{\omega\to\infty} \frac{\frac{\partial p}{\partial a}}{p} = -\frac{1}{1-a}$

For (ii):

$$\begin{aligned}
\frac{\frac{\partial p}{\partial b}}{p} &= \frac{\frac{\partial}{\partial b}\left(\frac{\gamma}{\mu(1-a)+b} + (1-a)\,s\right)}{p} = -\frac{\frac{\gamma}{(\mu(1-a)+b)^2}}{\frac{\gamma}{\mu(1-a)+b} + (1-a)\,\omega} \\
&= -\frac{\gamma}{\left(\mu\,(1-a)+b\right)\left(\gamma + (1-a)^2\omega\mu + \mu(1-a)b\right)}
\end{aligned}$$

hence $\frac{\frac{\partial p}{\partial b}}{p}$ is increasing in ω and $\lim_{\omega\to\infty} \frac{\partial p/\partial b}{p} = 0$. □

This simple framework makes precise that the effectiveness of the two policy levers depends on the efficiency of the monitor relative to the procurement officer. Because of this, offering the two jointly is either detrimental or inconsequential:

Proposition 5. *A joint increase in autonomy and incentives:*

(i) reduces prices by less than incentives alone when ω is low relative to h

(ii) converges to the effect of autonomy alone as $\omega \to \infty$.

Proof. Consider the combined treatment that changes autonomy by da and incentives by db. The effect of this is to change prices by

$$\begin{aligned} dp &= \frac{\partial p}{\partial a}da + \frac{\partial p}{\partial b}db + \frac{\partial^2 p}{\partial a \partial b}dadb \\ &= \left(\frac{\gamma\mu}{\left(\mu\left(1-a\right)+b\right)^2} - \omega\right)da - \frac{\gamma}{\left(\mu\left(1-a\right)+b\right)^2}db - \frac{2\gamma\mu}{\left(\mu\left(1-a\right)+b\right)^3}dadb \end{aligned}$$

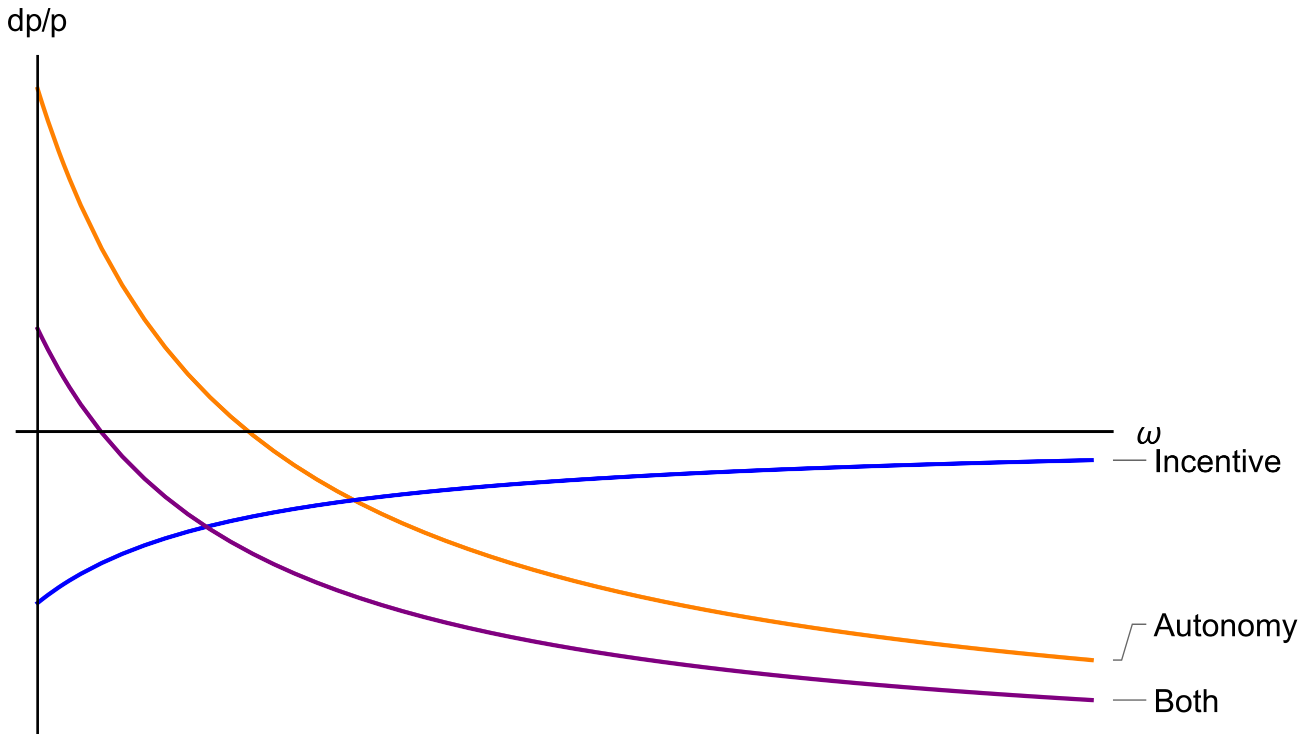
To see (i) compare the price change from the combined treatment to the price change resulting from a treatment that changes incentives by the same amount db but leaves autonomy unchanged. It is

$$\frac{da}{\left(\mu\left(1-a\right)+b\right)^3}\left[\gamma\mu\left(\mu\left(1-a\right)+b\right) - \omega\left(\mu\left(1-a\right)+b\right)^3 - 2\gamma\mu db\right]$$

which is negative as long as $\omega < \bar{\omega} - \frac{2\gamma\mu}{(\mu(1-a)+b)^3}db$ where $\bar{\omega}$ is as defined in the proof of proposition 4. (ii) follows from application of l'Hôpital's rule: $\lim_{\omega\to\infty} dp/p = 1/\left(1-a\right)$ which is the same as the limit of the autonomy treatment effect. □

The predictions of the model for the treatment effects and how they vary with the misalignment of the monitor ω are summarized graphically in figure C.1.

FIGURE C.1: MODEL PREDICTIONS OF HETEROGENEITY OF TREATMENT EFFECTS BY MONITOR TYPE



Notes: The figure shows the predictions our model in section C makes about how the treatment effects of our experiment will vary with the degree of misalignment of the monitor (ω) as described in propositions 4 and 5.

D Construction of Item Variety Measures

This appendix describes the methods we used to construct the item variety measures used throughout the empirical analysis. The idea behind the methods is to use data from the experiment's control group to construct measures in both treatment and control groups that allow us to hold constant all the features of the good that can affect its price in the control group. This poses two challenges. First, the set of attributes of each good may be large. Of these, only a subset is relevant for prices, and we want to avoid overfitting the data from the control group, so we want to reduce the dimensionality of the controls we use. Second, when using the control group data to construct measures of item variety in the treatment groups, the attributes used as inputs to these measures may not have common support. There may be attributes that occur in the treatment groups that never appear in the data from the control group. Our measures will predict how attributes that occur in the control group affect prices, but will not know how to deal with an attribute that only ever occurs in the treatment groups.

Our first three measures address these issues through manual grouping of attributes and using hedonic regressions to reduce the dimensionality of the measures. We begin by manually grouping attributes to ensure common support and avoid overfitting. Most of the attributes we use are categorical and so we group values. For values that occur less than three times in the control group or only in the treatment group, we either group them together with similar values (using contextual knowledge and extensive googling to find similar values) or if similar values are not available, set them to missing. Observations with all attributes missing after this cleaning are dropped. Ensuring that each group appears at least three times avoids overfitting, and ensuring that the groups are observed in both the control and treatment groups ensures common support. These groups then form the $\mathbf{X}_{igto}$ controls used in the hedonic regressions (1). Table D.1 illustrates the procedure. The first columns show the attributes in the raw data and the number of categories (for categorical variables) or the mean and standard deviation (for numerical variables) for each one. The second set of columns shows the same statistics for the data used for the hedonic regressions and the main analysis.

Our fourth, machine learning, measure develops a variant of a random forest algorithm to allow for non-linearities and interactions between attributes that the hedonic regression 1 rules out and also to perform the grouping of attributes' values in a data-driven way. For this we do much lighter cleaning of the data only harmonizing spellings. This can be seen in the third group of columns in table D.1, where the attributes tend to have a far greater number of categories. We then train a random forest algorithm for each item, averaging 500 trees to form predicted prices. The algorithm is trained only on the control group's data, so as in the case of the scalar and coarse measures of item variety, the predicted prices should be interpreted as a prediction of the price of the purchase had it been conducted by a PO in the control group.

After training each tree in the control group, the algorithm places each observation in the treatment groups into its corresponding leaf. It first places all treatment group observations that only have attributes that are sufficient to place it into a unique leaf in the tree. Then, for observations that have an attribute that prevents it from being placed into a leaf, the algorithm selects all leaves the observation could be placed into given the attributes that *can* be used, and then for each attribute that cannot be used, replaces that attribute with the category in the same treatment group with the closest average, but that does appear in the control group. Once every observation is placed into a leaf, the average price amongst control group observations in the leaf is then that tree's predicted price. Averaging the 500 trees gives us our machine learning measure of item variety.

Finally, table D.2 shows that the main analysis is also robust to including observations which we deemed to have insufficient attributes to construct our measures of item variety. Column 1 presents results from running our difference in difference specification to estimate the impacts of the autonomy and combined treatments. These results are comparable to those in column 1 of table E.2. Column 2 presents results from our baseline specification using only data from year 2 of the experiment. These results are comparable to those in column 3 of table 2, but, consistent with them having insufficient information to precisely measure item variety and this introducing noise into the estimation, they are less precise.

TABLE D.1: POPS DATA CLEANING

Item	Attributes	Raw Data: mean (s.d.) / #categories	Raw Data: # missing	Regression Data: mean (s.d.) / #categories	Regression Data: # missing	Machine Learning Data: mean (s.d.) / #categories	Machine Learning Data: # missing
Pencil	Brand	21 categories	272	8 categories	187	19 categories	156
	Grade	26 categories	279	13 categories	175	25 categories	159
	Type	8 categories	156	5 categories	54	5 categories	46
	With Rubber?	2 categories	281	2 categories	177	2 categories	164
	Unit Price. mean (s.d.)	10.81 (14.91)		10.51 (14.42)		9.80 (11.31)	
	# Purchasing PBs	311		275		253	
	# Observations	612		476		475	
Ice Block	Unit Price. mean (s.d.)	0.01 (0.02)		0.01 (0.01)		0.01 (0.01)	
	# Purchasing PBs	321		304		304	
	# Observations	680		638		638	
Wiper	Brand	13 categories	388	4 categories	173	12 categories	152
	Country of Origin	3 categories	331	2 categories	98	2 categories	98
	Handle Length	8 categories	381	5 categories	141	5 categories	141
	Handle Material	5 categories	304	4 categories	77	4 categories	77
	Wiper Material	7 categories	314	3 categories	88	3 categories	87
	Unit Price. mean (s.d.)	271.42 (125.82)		264.13 (115.92)		264.13 (115.92)	
	# Purchasing PBs	401		296		296	
	# Observations	753		484		484	

Continued on next page

Table D.1: POPS Data Cleaning

Item	Attributes	Raw Data		Regression Data		Machine Learning Data	
		mean (s.d.) / #categories	# missing	mean (s.d.) / #categories	# missing	mean (s.d.) / #categories	# missing
Calculator	Brand & Model	7 categories	150	12 categories	49	22 categories	44
	Number of Digits	6 categories	205	4 categories		4 categories	
	Type	5 categories	185	4 categories	76	4 categories	77
	Unit Price. mean (s.d.)	271.42 (125.82)		796.24 (350.34)		795.93 (350.05)	
	# Purchasing PBs	401		326		326	
	# Observations	616		486		487	
Coal	Unit Price. mean (s.d.)	0.08 (0.26)		0.06 (0.02)		0.06 (0.02)	
	# Purchasing PBs	384		362		362	
	# Observations	685		650		650	
Staples	Brand	19 categories	69	8 categories	59	19 categories	36
	Size	27 categories	60	6 categories	26	5 categories	26
	Unit Price. mean (s.d.)	0.14 (0.43)		0.11 (0.20)		0.11 (0.20)	
	# Purchasing PBs	334		288		288	
	# Observations	551		465		465	

Continued on next page

TABLE D.1: POPS DATA CLEANING

Item	Attributes	Raw Data mean (s.d.) / #categories	Raw Data # missing	Regression Data mean (s.d.) / #categories	Regression Data # missing	Machine Learning Data mean (s.d.) / #categories	Machine Learning Data # missing
Lock	Brand & Model	18 categories	508	4 categories	270	9 categories	231
	Country of Origin	5 categories	384	2 categories	117	2 categories	119
	Digital?	2 categories	526	2 categories	245	2 categories	247
	Fitting Charges?	2 categories	514	2 categories	235	2 categories	237
	Size	27 categories	60	6 categories	26	5 categories	26
	Material	8 categories	512	4 categories	233	4 categories	235
	Type	20 categories	440	7 categories	166	13 categories	160
	Unit Price. mean (s.d.)	315.94 (340.11)		282.89 (235.49)		282.56 (235.21)	
	# Purchasing PBs	404		318		319	
	# Observations	965		652		654	
Stamp Pad	Brand	19 categories	262	10 categories	77	18 categories	64
	Color	8 categories	281	5 categories	86	6 categories	86
	Size	22 categories	317	8 categories	125	8 categories	125
	With Ink?	3 categories	266	2 categories	81	2 categories	81
	Unit Price. mean (s.d.)	85.92 (50.40)		82.72 (44.05)		82.98 (43.92)	
	# Purchasing PBs	430		352		352	
	# Observations	771		545		543	

Continued on next page

TABLE D.1: POPS DATA CLEANING

Item	**Attributes**	**Raw Data** mean (s.d.) / #categories	**Raw Data** # missing	**Regression Data** mean (s.d.) / #categories	**Regression Data** # missing	**Machine Learning Data** mean (s.d.) / #categories	**Machine Learning Data** # missing
Duster	Material	9 categories	261	6 categories	37	7 categories	37
	Size	52 categories	437	17 categories	195	18 categories	193
	Type	9 categories	343	4 categories	116	4 categories	116
	With Handle?	2 categories	435	2 categories	196	2 categories	196
	Unit Price. mean (s.d.)	66.31 (76.83)		65.13 (71.31)		65.13 (71.31)	
	# Purchasing PBs	386		290		290	
	# Observations	722		456		456	
Floor Cleaner	Acid Cleaner	7 categories	376	4 categories	242	4 categories	235
	Brand	38 categories	348	16 categories	258	30 categories	216
	Environmentally Friendly	2 categories	286	2 categories	168	2 categories	169
	Make	6 categories	307	4 categories	180	6 categories	177
	Scented	2 categories	230	2 categories	116	2 categories	117
	State	8 categories	225	3 categories	103	3 categories	104
	Unit Price. mean (s.d.)	0.27 (0.94)		0.19 (0.30)		0.19 (0.30)	
	# Purchasing PBs	458		377		377	
	# Observations	1162		945		946	

Continued on next page

Table D.1: POPS Data Cleaning

Item	Attributes	Raw Data mean (s.d.) / #categories	Raw Data # missing	Regression Data mean (s.d.) / #categories	Regression Data # missing	Machine Learning Data mean (s.d.) / #categories	Machine Learning Data # missing
File Cover	Brand	20 categories	399	5 categories	306	18 categories	286
	With Clip	2 categories	662	2 categories	258	2 categories	259
	Country of Origin	6 categories	379	4 categories	265	3 categories	266
	Cover Material	22 categories	244	11 categories	150	13 categories	151
	Customized Printing	5 categories	328	4 categories	228	3 categories	229
	File Type	28 categories	138	14 categories	61	22 categories	58
	Size	27 categories	414	3 categories	290	3 categories	291
	Unit Price. mean (s.d.)	53.11 (95.41)		47.62 (75.07)		47.56 (75.02)	
	# Purchasing PBs	391		312		313	
	# Observations	775		583		584	
Sign Board / Banner	Frame Type	7 categories	667	3 categories	586	5 categories	586
	Material	11 categories	445	7 categories	391	10 categories	391
	Number of Colors	6 categories	723	2.8 (1.23)	643	2.8 (1.23)	643
	Number of Rings	12 categories	692	4.4 (4.05)	1055	4.4 (4.05)	1055
	Print on Both Sides	3 categories	625	2 categories	550	2 categories	551
	Area	85 categories	732	44.2 (355.64)	644	44.2 (355.64)	644
	With Rope	2 categories	598	2 categories	523	2 categories	523
	With Stand	2 categories	598	2 categories	519	2 categories	519
	With Stick	2 categories	590	2 categories	511	2 categories	511

Continued on next page

Table D.1: POPS Data Cleaning

Item	Attributes	Raw Data: mean (s.d.) / #categories	Raw Data: # missing	Regression Data: mean (s.d.) / #categories	Regression Data: # missing	Machine Learning Data: mean (s.d.) / #categories	Machine Learning Data: # missing
	Unit Price. mean (s.d.)	1,262.06 (1,881.76)		1,170.37 (1,557.29)		1,170.37 (1,557.29)	
	# Purchasing PBs	442		402		402	
	# Observations	1391		1256		1256	
Stapler	Brand & Model	60 categories	584	15 categories	176	28 categories	149
	Size	9 categories	566	4 categories	123	4 categories	141
	Unit Price. mean (s.d.)	587.33 (816.28)		507.08 (621.07)		504.22 (614.41)	
	# Purchasing PBs	539		364		372	
	# Observations	1024		549		567	
Photocopying	Color	2 categories	1119	2 categories	307	2 categories	307
	Double-sided	3 categories	1248	3 categories	395	3 categories	395
	On Generator Power	3 categories	1175	3 categories	370	3 categories	370
	Paper Quality	9 categories	1693	3 categories	831	7 categories	831
	Size	19 categories	1043	3 categories	221	12 categories	215
	With Binding	4 categories	1585	3 categories	725	3 categories	725
	Unit Price. mean (s.d.)	3.33 (7.65)		2.69 (2.76)		2.69 (2.76)	
	# Purchasing PBs	470		401		401	
	# Observations	3185		2249		2249	
Toner	Brand & Model	180 categories	1280	57 categories	581	31 categories	581
	Refill or New	7 categories	935	5 categories	241	5 categories	241

Continued on next page

Table D.1: POPS Data Cleaning

Item	Attributes	Raw Data mean (s.d.) / #categories	Raw Data # missing	Regression Data mean (s.d.) / #categories	Regression Data # missing	Machine Learning Data mean (s.d.) / #categories	Machine Learning Data # missing
	Unit Price. mean (s.d.)	4,630.16 (4,257.79)		4,449.26 (3,873.94)		4,449.26 (3,873.94)	
	# Purchasing PBs	505		449		449	
	# Observations	3814		2980		2980	
Envelope	Material	12 categories	789	7 categories	417	10 categories	417
	Printed	5 categories	983	4 categories	583	4 categories	583
	Area	5 categories	983	4 categories	583	4 categories	583
	With Zip	2 categories	1112	2 categories	726	2 categories	727
	Unit Price. mean (s.d.)	9.31 (32.16)		6.40 (14.18)		6.38 (14.16)	
	# Purchasing PBs	512		427		427	
	# Observations	1891		1433		1438	
Soap	Antiseptic	2 categories	690	2 categories	418	2 categories	420
	Brand	36 categories	436	20 categories	209	30 categories	192
	State	3 categories	419	3 categories	181	3 categories	183
	Type	19 categories	544	9 categories	314	11 categories	318
	Bar Size	67 categories	0	198.1 (137.86)	0	198.0 (137.75)	0
	Bottle Size	67 categories	0	0.9 (0.71)	0	0.9 (0.71)	0
	Packet Size	67 categories	0	1072.1 (2461.58)	0	1072.0 (2459.27)	0

Continued on next page

Table D.1: POPS Data Cleaning

Item	Attributes	Raw Data mean (s.d.) / #categories	Raw Data # missing	Regression Data mean (s.d.) / #categories	Regression Data # missing	Machine Learning Data mean (s.d.) / #categories	Machine Learning Data # missing
	Unit Price. mean (s.d.)	3.73 (17.96)		2.17 (11.14)		2.17 (11.12)	
	# Purchasing PBs	518		446		447	
	# Observations	1476		1155		1158	
Light Bulb	Brand	53 categories	959	12 categories	434	31 categories	386
	Type	28 categories	772	9 categories	224	22 categories	209
	Wattage	47 categories	814	12 categories	232	35.4 (65.15)	252
	With Fitting	3 categories	1505	2 categories	862	2 categories	882
	With Fixture	3 categories	1463	2 categories	818	2 categories	838
	Unit Price. mean (s.d.)	697.49 (1,142.68)		541.53 (747.52)		563.52 (782.47)	
	# Purchasing PBs	530		446		446	
	# Observations	1818		1173		1193	
Broom	Brand	8 categories	846	4 categories	380	8 categories	369
	Handle Length	10 categories	815	3.1 (1.57)	878	3.1 (1.57)	878
	Handle Material	4 categories	838	4 categories	351	4 categories	351
	Type	23 categories	588	10 categories	139	15 categories	121
	Unit Price. mean (s.d.)	79.90 (108.92)		76.36 (102.71)		76.36 (102.71)	
	# Purchasing PBs	586		455		455	
	# Observations	1702		1159		1159	
Newspaper	Name	57 categories	2129	23 categories	0	29 categories	0

Continued on next page

Table D.1: POPS Data Cleaning

Item	Attributes	Raw Data: mean (s.d.) / #categories	Raw Data: # missing	Regression Data: mean (s.d.) / #categories	Regression Data: # missing	Machine Learning Data: mean (s.d.) / #categories	Machine Learning Data: # missing
	Unit Price. mean (s.d.)	14.74 (6.09)		14.29 (3.72)		14.29 (3.72)	
	# Purchasing PBs	717		617		618	
	Unit Price. mean (s.d.)	14.74 (6.09)		14.29 (3.72)		14.29 (3.72)	
	# Purchasing PBs	717		617		618	
	# Observations	9400		6647		6683	
Register	Binding	15 categories	2917	13 categories	1633	10 categories	1635
	Brand	54 categories	3209	19 categories	1979	49 categories	1920
	Colored Pages	6 categories	2933	2 categories	1675	2 categories	1677
	Customized Printing	3 categories	3011	2 categories	1732	2 categories	1734
	Number of Pages	80 categories	2939	185.1 (169.65)	1641	185.1 (169.65)	1643
	Page Size	82 categories	2874	26 categories	1552	51 categories	1554
	Page Weight	14 categories	4456	12 categories	2602	14 categories	2604
	Type	114 categories	1776	28 categories	523	44 categories	525
	Unit Price. mean (s.d.)	14.74 (6.09)		314.93 (239.41)		314.84 (239.38)	
	# Purchasing PBs	717		717		718	
	# Observations	5176		3705		3707	

Continued on next page

Table D.1: POPS Data Cleaning

Item	**Attributes**	**Raw Data** mean (s.d.) / #categories	**Raw Data** # missing	**Regression Data** mean (s.d.) / #categories	**Regression Data** # missing	**Machine Learning Data** mean (s.d.) / #categories	**Machine Learning Data** # missing
Printer Paper	Brand	33 categories	1127	14 categories	693	31 categories	638
	Colored Pages	3 categories	1014	2 categories	531	2 categories	532
	Page Size	21 categories	1123	7 categories	547	15 categories	547
	Page Weight	25 categories	898	13 categories	360	77.54 (5.99)	361
	Unit Price. mean (s.d.)	1.30 (1.49)		1.19 (0.28)		1.19 (0.28)	
	# Purchasing PBs	837		746		746	
	# Observations	4570		3842		3843	
Pen	Color	15 categories	1579	11 categories	911	8 categories	912
	Model	59 categories	1560	29 categories	916	30 categories	887
	Type	15 categories	978	8 categories	349	9 categories	350
	Thickness	23 categories	2188	1.1 (1.04)	1443	1.1 (1.04)	1444
	Unit Price. mean (s.d.)	49.10 (126.38)		40.26 (58.98)		40.27 (58.98)	
	# Purchasing PBs	814		719		719	
	# Observations	4298		3386		3387	
Towel	Size	24 categories	517	1137.6 (446.45)	334	1137.6 (446.45)	334
	Towel Material	3 categories	283	2 categories	109	2 categories	109
	Type	7 categories	198	4 categories	32	4 categories	32

Continued on next page

Table D.1: POPS Data Cleaning

Item	Attributes	Raw Data: mean (s.d.) / #categories	Raw Data: # missing	Regression Data: mean (s.d.) / #categories	Regression Data: # missing	Machine Learning Data: mean (s.d.) / #categories	Machine Learning Data: # missing
	Unit Price. mean (s.d.)	458.19 (225.20)		469.73 (206.82)		469.73 (206.82)	
	# Purchasing PBs	362		273		273	
	# Observations	617		427		427	
Pipe	Diameter	60 categories	365	2.0 (3.59)	207	1.9 (3.58)	207
	Manufacturer	32 categories	414	10 categories	273	22 categories	243
	Material	3 categories	283	5 categories	94	13 categories	81
	Size	62 categories	441	15 categories	316	607.5 (1068.00)	316
	Type	41 categories	326	39 categories	162	30 categories	162
	Unit Price. mean (s.d.)	2.30 (8.63)		1.87 (6.26)		1.87 (6.26)	
	# Purchasing PBs	372		319		319	
	# Observations	807		609		610	
TOTAL	# Observations	49,461		36,950		37,039	

Table D.2: Robustness of Price Effects to Including POPS Observations With Insufficient Attributes

	(1) DiD	(2) DiD	(3) Year 2	(4) Year 2
Autonomy			-0.063	-0.050
			(0.044)	(0.031)
			[0.209]	[0.165]
Incentives			-0.000	0.004
			(0.042)	(0.029)
			[0.993]	[0.909]
Both			-0.036	-0.047
			(0.042)	(0.031)
			[0.466]	[0.193]
Autonomy $\times$ Year 2	-0.078	-0.071		
	(0.050)	(0.040)		
	[0.102]	[0.046]		
Both $\times$ Year 2	-0.082	-0.084		
	(0.051)	(0.041)		
	[0.075]	[0.028]		
Year 2	-0.001	0.019		
	(0.042)	(0.032)		
Item Variety Control	None	Attribs	None	Attribs
p(All = 0)	0.095	0.038	0.545	0.262
p(Autonomy = Incentives)			0.212	0.112
p(Autonomy = Both)	0.101	0.747	0.605	0.921
p(Incentives = Both)			0.441	0.133
Observations	25,254	25,254	12,933	12,933

Notes: The table shows estimates of the treatment effects of the experiments on log unit prices. The sample used extends our main analysis sample to also include observations from POPS that were dropped because they contained insufficient detail on the attributes of the items being purchased. Column 1 presents results from running our difference in difference specification to estimate the impacts of the autonomy and combined treatments. These results are comparable to those in column 1 of table E.2. Column 2 presents results from our baseline specification using only data from year 2 of the experiment. These results are comparable to those in column 3 of table 2.

E Additional Results: Average Treatment Effects

This appendix presents additional results from the analysis of average treatment effects.

Order Sizes as Bad Controls In section 5.2 we discuss a potential composition effect that would arise if the experiment directly affected the varieties of items procurement officers purchase and these are then included as controls in price regressions. A similar concern arises with the size of the order since we control for the size of the order in order to capture bulk discounts, but the size of the orders may be directly affected by the experiment.

However, as with the variety of items purchased, we see no effects of the experiment on the sizes of orders. Table E.1 shows estimates of overall treatment effects of the three treatments on the log quantity purchased in each order. The table shows estimates of equation:

$$q_{igto} = \alpha + \sum_{k=1}^{3} \eta_k \text{Treatment}_o^k + \mathbf{X}_{igto}\beta + \delta_s + \gamma_g + \varepsilon_{igto}$$

where q_{igto} is the log quantity in purchase i of good g at time t by office o in columns 1–5, the log "value" of the order (log quantity plus log "price") priced using the scalar control in column (6), and the log value priced using the ML control in column (7). Treatment$_o^k$ indicates the three treatment groups; δ_s and γ_g are stratum and good fixed effects, respectively; and $\mathbf{X}_{igto}$ are purchase-specific controls. In column (2) these controls include all item attributes, in column (3) the scalar item variety measure, in column (4) the coarse item variety, and in column (5) the machine-learning item variety measure. We weight regressions by expenditure shares in the control group so that treatment effects can be interpreted as effects on expenditure, and the residual term ε_{igto} is clustered at the cost center level. Below each coefficient we report standard errors clustered by cost center in parentheses, and p-values from randomization inference tests of the hypothesis that the treatment has no effect on any office in square brackets. Across the board, we do not see any impact of the experiment on order sizes.

Difference in Differences Analysis to Deal with Composition of Purchases As an alternative way of controlling for the composition of purchases, we exploit the data from year 1 of the project to estimate treatment effects of the introduction of autonomy through a difference in differences approach. This allows us to control for office fixed effects so that we exploit only within-office changes, allowing us to hold constant the component of the composition effect $\mathbb{E}\left[p\left(0,1\right)|H\right]$ that comes from office-level variation in the types of items demanded.

Specifically, we estimate difference in differences specifications of the form

$$y_{igto} = \alpha + \sum_{k=1}^{3} \eta_k \text{Treatment}_o^k \times \text{Year2}_t + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \gamma_g + \delta_t + \lambda_o + \varepsilon_{igto}$$

where y_{igto} is the outcome of interest. Appendix table E.2 shows the results. In columns 1–3 the outcome is the scalar (column 1), coarse (column 2) or machine learning (column 3) measure of good variety, while in columns 4–8 it is the log unit price. Treatment_o^k indicates the three treatment groups (though we only report coefficients for the autonomy and both treatments since the incentives treatment was already in place in year 1); Year2_t indicates purchases in year 2; $\mathbf{X}_{igto}$ are purchase-level controls; q_{igto} is the quantity purchased; γ_g, δ_t and λ_o are good-, year- and office- fixed effects, respectively; and ε_{igto} are residuals clustered by office. Column 5 controls for the full vector of item attributes, column 6 for the scalar item variety measure, column 7 for the coarse item variety measure, and column 8 uses the machine learning measure of item variety. Below each coefficient we report standard errors clustered by office in parentheses and the p-values from randomization inference on the hypothesis that the treatment effect is zero for all offices.The table shows again that there are no discernible effects on the varieties of the goods being purchased, and that the treatment effects on prices are, if anything, slightly larger than in table 2.

Experience Effects We do not find evidence that the experiment had delayed effects due to procurement officers learning over time that the treatments were effective. In table E.3 we reestimate the effects of the treatments, interacting them with the time at which the purchase was made and the order in which the purchases were made:

$$y_{igto} = \alpha + \sum_{k=1}^{3} \left(\eta_k \text{Treatment}_o^k + \kappa_k \text{Treatment}_o^k \times \text{Time}_{ito} \right) + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$$

where Treatment_o^k are dummies for office o being in treatment k; Time_{ito} is a measures of time, calendar time (scaled to be 0 at the beginning of the fiscal year and 1 at the end of the year) and/or the order of the purchase made by the office (scaled to be between 0 and 1); $\mathbf{X}_{igto}$ is a vector of controls; q_{igto} is the quantity purchased, δ_s and γ_g are strata and good fixed effects, respectively, and ε_{igto} are residuals clustered by office. Columns 1–6 estimate dynamic treatment effects on the variety purchased using the scalar measure (columns 1–3) and coarse measure (columns 4–6) described in section 5.1. Columns 7–18 estimate dynamic treatment effects on log unit prices paid, not controlling for the variety purchased (columns 7–9), or controlling for the variety purchased using the full vector of good attributes (columns 10–12), the scalar variety measure (13–15), or the coarse variety measure (16-18).

We find no evidence that the treatments had any dynamic effect on procurement performance. The estimated treatment effects at the beginning of the year are indistinguishable from the overall effects in table 2, and all the interaction terms are indistinguishable from zero at 5% significance. This also suggests that POs did not try to game the incentive treatment by reducing prices early on to win an interim prize and then recouping their losses later in the year.

Generic Budget Share While we do not find evidence of experience effects, we *do*, however, find that the experiment had larger effects on offices for whom generic goods form a larger share of their annual budget, as shown in table E.4. These are offices where purchasing generics is a larger part of the job of the procurement officer and so the treatments have a bigger impact as one would expect.

Quantities Demanded The results in table 2 lead us to conclude that the treatments lowered prices paid without affecting the varieties of the items being purchased. We might naturally expect that if the prices at which goods can be procured go down, offices react by increasing demand for goods. On the other hand, since the demand for goods is coming from end users, while the procurement officer simply fulfills their orders, we might not expect these lower prices to pass through to end users' demand.

To investigate the impacts of the treatments on the quantities purchased and expenditure, we value each purchase using the counterfactual prices we estimate each purchase would have been made at had it been made by an office in the control group–the scalar variety measure. That is, for each purchase, the counterfactual expenditure is $e_{igto} = \exp\left(v_{igto} + q_{igto}\right)$ where v_{igto} is the scalar good variety measure, and q_{igto} is the log number of units purchased. We then aggregate the data to the good-month-office level and estimate good-specific treatment effects by multivariate regression with the following specification for each item

$$e_{gto} = \sum_{k=1}^{3} \eta_{kg} \text{Treatment}_o^k + \gamma_s + \xi_t + \varepsilon_{gto} \tag{E.1}$$

where e_{gto} is the quantity purchased of good g in month t by office o; the η_{kg} are good-specific treatment effects; γ_s and ξ_t are stratum and month fixed effects respectively; and ε_{gto} are residuals clustered by office.

Table E.5 shows the results. For each good, we display the estimated η_{kg} coefficients and their standard errors clustered by office, as well as the F-statistic for the hypothesis that all three η_{kg}s are equal to zero and its p-value in square brackets. We also display F-statistics for the hypothesis that each treatment has zero effect on any item, and the F-statistic on the hypothesis that none of the treatments affect any of the items.

Of the 75 estimated η_{kg} treatment effects, only two are statistically significant at the 5% level, consistent with what would be expected purely by chance, and for all but three items, we fail to reject the hypothesis that all three treatments have no effect. Similarly, we cannot reject the hypotheses that each treatment affects none of the items or the hypothesis that no treatment affects any item. As a result, we conclude that there is no evidence that any of the treatments affected the composition of offices' expenditure or the overall amount they purchase. Of course, this inelastic demand could be because end users truly have inelastic demand (for example due to capacity constraints) or because of agency issues within the office whereby price reductions achieved by the procurement officer are not passed through to end users, however distinguishing between these two remains an open question.

Procurement Timing A final margin along which procurement officers might respond is by changing the timing of their procurement. If there is predictable seasonality in prices, the incentives treatment might cause procurement officers to shift purchases into lower-price times of the year. If monitoring by the AG leads to delays in procurement, we might expect the autonomy treatment to permit procurement officers to make purchases more quickly. On the other hand, table E.5 suggests offices' demand is inelastic with respect to price, and so if the timing of demand is also inelastic (e.g. goods are required to coincide with the start of the school year) then we might not expect our experiment to affect the timing of procurement.

Figure E.1 shows estimates of treatment effects on the timing of deliveries and expenditure. The estimates are from seemingly unrelated regressions of the form

$$\mathbf{1}\left\{\text{Month}_i = m\right\} = \alpha + \beta_A \text{Autonomy}_i + \beta_I \text{Incentives}_i + \beta_B \text{Both}_i + \gamma_g + \gamma_s + \varepsilon_i$$

where γ_g are good fixed effects, γ_s are randomization strata fixed effects, and ε_i are residuals clustered by office. The figures show the 95% confidence intervals of the estimated β_A, β_I and β_B with p-values of χ^2 tests of the hypothesis that each treatment's effect is 0 in all months, and the hypothesis that all treatments have no effect in all months. The 95% confidence intervals include zero for all months and treatments except the autonomy treatment in December. Moreover, we are unable to reject the hypotheses that each treatment has zero effect in all months or the hypothesis that none of the treatments affect the probability of delivery in any month.

TABLE E.1: ORDER SIZES ARE UNAFFECTED

	Quantity					CF Value	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Autonomy	-0.044	-0.004	-0.026	-0.045	-0.042	-0.028	-0.042
	(0.068)	(0.060)	(0.053)	(0.068)	(0.064)	(0.054)	(0.065)
	[0.537]	[0.948]	[0.629]	[0.524]	[0.559]	[0.612]	[0.557]
Incentives	0.029	0.079	0.036	0.039	0.042	0.035	0.039
	(0.070)	(0.064)	(0.056)	(0.070)	(0.064)	(0.057)	(0.065)
	[0.702]	[0.250]	[0.544]	[0.614]	[0.563]	[0.555]	[0.595]
Both	-0.096	-0.059	-0.055	-0.091	-0.085	-0.060	-0.088
	(0.071)	(0.066)	(0.060)	(0.070)	(0.067)	(0.060)	(0.067)
	[0.207]	[0.418]	[0.415]	[0.239]	[0.226]	[0.370]	[0.217]
Item Variety Control	None	Attribs	Scalar	Coarse	ML	Scalar	ML
p(All = 0)	0.362	0.306	0.559	0.365	0.357	0.535	0.360
Observations	11,771	11,771	11,771	11,771	11,771	11,771	11,771

Notes: The table shows estimates of overall treatment effects of the three treatments on the log quantity purchased in each order. The table shows estimates of equation:

$$q_{igto} = \alpha + \sum_{k=1}^{3} \eta_k \text{Treatment}_o^k + \mathbf{X}_{igto}\beta + \delta_s + \gamma_g + \varepsilon_{igto}$$

where q_{igto} is the log quantity in purchase i of good g at time t by office o in columns 1–5, the log "value" of the order (log quantity plus log "price") priced using the scalar control in column (6), and the log value priced using the ML control in column (7). Treatment$_o^k$ indicates the three treatment groups; δ_s and γ_g are stratum and good fixed effects, respectively; and $\mathbf{X}_{igto}$ are purchase-specific controls. In column (2) these controls include all item attributes, in column (3) the scalar item variety measure, in column (4) the coarse item variety, and in column (5) the machine-learning item variety measure. We weight regressions by expenditure shares in the control group so that treatment effects can be interpreted as effects on expenditure, and the residual term ε_{igto} is clustered at the cost center level. Below each coefficient we report standard errors clustered by cost center in parentheses, and p-values from randomization inference tests of the hypothesis that the treatment has no effect on any office in square brackets.

TABLE E.2: DIFFERENCE IN DIFFERENCES TREATMENT EFFECTS

	Variety			Unit Price				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Autonomy × Year 2	-0.009	0.015	-0.013	-0.128	-0.130	-0.122	-0.132	-0.127
	(0.023)	(0.029)	(0.012)	(0.047)	(0.043)	(0.043)	(0.045)	(0.047)
	[0.720]	[0.564]	[0.283]	[0.010]	[0.003]	[0.007]	[0.004]	[0.003]
Both × Year 2	0.019	0.049	0.006	-0.098	-0.117	-0.112	-0.102	-0.099
	(0.027)	(0.033)	(0.012)	(0.050)	(0.042)	(0.043)	(0.045)	(0.049)
	[0.505]	[0.131]	[0.608]	[0.045]	[0.005]	[0.016]	[0.025]	[0.045]
Item Variety Measure	Scalar	Coarse	ML	None	Attribs	Scalar	Coarse	ML
p(All = 0)	0.736	0.478	0.408	0.053	0.008	0.021	0.031	0.048
p(Autonomy = Both)	0.238	0.270	0.097	0.542	0.741	0.831	0.535	0.578
Observations	21,183	21,183	21,182	21,183	21,183	21,183	21,183	21,182

Notes: The table shows difference in differences estimates of the treatment effect of the introduction of the autonomy treatment in year 2 of the experiment. The estimates in columns 1–3 are of regressions of the form

$$y_{igto} = \alpha + \sum_{k=1}^{3} \eta_k \text{Treatment}_o^k \times \text{Year2}_t + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \gamma_g + \delta_t + \lambda_o + \varepsilon_{igto}$$

where y_{igto} is the outcome of interest. In columns 1–3 it is the scalar (column 1), coarse (column 2) or machine learning (column 3) measure of good variety, while in columns 4–8 it is the log unit price. Treatment$_o^k$ indicates the three treatment groups (though we only report coefficients for the autonomy and both treatments since the incentives treatment was already in place in year 1); Year2$_t$ indicates purchases in year 2; $\mathbf{X}_{igto}$ are purchase-level controls; q_{igto} is the quantity purchased; γ_g, δ_t and λ_o are good-, year- and office- fixed effects, respectively; and ε_{igto} are residuals clustered by office. Column 5 controls for the full vector of item attributes, column 6 for the scalar item variety measure, column 7 for the coarse item variety measure, and column 8 uses the machine learning measure of item variety. Below each coefficient we report standard errors clustered by office in parentheses and the p-values from randomization inference on the hypothesis that the treatment effect is zero for all offices.

TABLE E.3: DYNAMIC TREATMENT EFFECTS ON PRICES PAID

	Variety						Unit Price											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Autonomy	-0.028	-0.034	-0.037	-0.034	-0.056	-0.053	-0.148	-0.155	-0.159	-0.149	-0.139	-0.150	-0.121	-0.127	-0.129	-0.145	-0.143	-0.151
	(0.042)	(0.043)	(0.044)	(0.041)	(0.037)	(0.039)	(0.057)	(0.053)	(0.055)	(0.050)	(0.049)	(0.051)	(0.048)	(0.046)	(0.048)	(0.049)	(0.045)	(0.047)
	[0.783]	[0.662]	[0.700]	[0.545]	[0.308]	[0.395]	[0.130]	[0.075]	[0.108]	[0.048]	[0.042]	[0.048]	[0.049]	[0.020]	[0.035]	[0.062]	[0.033]	[0.046]
Incentives	-0.052	-0.036	-0.051	-0.023	-0.023	-0.031	-0.078	-0.078	-0.088	-0.038	-0.037	-0.039	-0.013	-0.038	-0.027	-0.050	-0.052	-0.055
	(0.046)	(0.043)	(0.045)	(0.037)	(0.035)	(0.038)	(0.073)	(0.052)	(0.064)	(0.055)	(0.042)	(0.049)	(0.077)	(0.051)	(0.064)	(0.064)	(0.047)	(0.056)
	[0.622]	[0.706]	[0.641]	[0.596]	[0.612]	[0.539]	[0.510]	[0.500]	[0.534]	[0.622]	[0.591]	[0.640]	[0.860]	[0.500]	[0.684]	[0.611]	[0.579]	[0.635]
Both	-0.063	-0.033	-0.054	0.075	0.069	0.073	-0.170	-0.164	-0.178	-0.129	-0.140	-0.142	-0.080	-0.079	-0.080	-0.184	-0.176	-0.190
	(0.041)	(0.046)	(0.044)	(0.045)	(0.043)	(0.047)	(0.057)	(0.053)	(0.055)	(0.043)	(0.044)	(0.044)	(0.051)	(0.049)	(0.051)	(0.056)	(0.053)	(0.054)
	[0.436]	[0.730]	[0.594]	[0.296]	[0.351]	[0.375]	[0.083]	[0.067]	[0.063]	[0.046]	[0.021]	[0.030]	[0.171]	[0.158]	[0.183]	[0.039]	[0.025]	[0.030]
Autonomy × Time	0.078		0.024	0.078		-0.034	0.113		0.045	0.113		0.102	0.072		0.022	0.113		0.079
	(0.055)		(0.086)	(0.056)		(0.090)	(0.070)		(0.128)	(0.059)		(0.095)	(0.061)		(0.100)	(0.061)		(0.114)
	[0.654]		[0.856]	[0.423]		[0.699]	[0.513]		[0.740]	[0.317]		[0.362]	[0.333]		[0.833]	[0.414]		[0.520]
Incentives × Time	0.105		0.113	0.086		0.054	0.112		0.071	0.024		0.018	-0.015		-0.078	0.054		0.029
	(0.057)		(0.089)	(0.050)		(0.072)	(0.103)		(0.165)	(0.075)		(0.128)	(0.111)		(0.177)	(0.089)		(0.150)
	[0.523]		[0.349]	[0.207]		[0.434]	[0.639]		[0.664]	[0.844]		[0.905]	[0.897]		[0.716]	[0.777]		[0.853]
Both × Times	0.179		0.217	-0.029		-0.042	0.180		0.142	0.084		0.026	0.016		0.014	0.176		0.147
	(0.056)		(0.104)	(0.070)		(0.077)	(0.078)		(0.145)	(0.057)		(0.118)	(0.075)		(0.124)	(0.077)		(0.145)
	[0.096]		[0.093]	[0.795]		[0.588]	[0.377]		[0.417]	[0.493]		[0.836]	[0.842]		[0.917]	[0.289]		[0.401]
Autonomy × Order		0.095	0.075		0.124	0.154		0.131	0.093		0.102	0.014		0.088	0.069		0.115	0.047
		(0.062)	(0.097)		(0.056)	(0.095)		(0.073)	(0.138)		(0.066)	(0.106)		(0.066)	(0.111)		(0.064)	(0.122)
		[0.497]	[0.412]		[0.179]	[0.135]		[0.613]	[0.615]		[0.467]	[0.905]		[0.342]	[0.593]		[0.513]	[0.741]
Incentives × Order		0.079	-0.012		0.092	0.048		0.118	0.061		0.022	0.008		0.030	0.092		0.059	0.036
		(0.055)	(0.087)		(0.048)	(0.072)		(0.070)	(0.131)		(0.056)	(0.109)		(0.069)	(0.134)		(0.060)	(0.121)
		[0.619]	[0.874]		[0.203]	[0.578]		[0.696]	[0.742]		[0.927]	[0.942]		[0.747]	[0.611]		[0.848]	[0.803]
Both × Order		0.128	-0.056		-0.019	0.016		0.173	0.053		0.105	0.083		0.014	0.002		0.165	0.041
		(0.066)	(0.118)		(0.067)	(0.062)		(0.071)	(0.143)		(0.061)	(0.129)		(0.073)	(0.124)		(0.070)	(0.142)
		[0.439]	[0.626]		[0.924]	[0.853]		[0.386]	[0.738]		[0.422]	[0.566]		[0.885]	[0.987]		[0.237]	[0.766]
Item Variety Control	Scalar	Scalar	Scalar	Coarse	Coarse	Coarse	None	None	None	Attribs	Attribs	Attribs	Scalar	Scalar	Scalar	Coarse	Coarse	Coarse
p(All = 0)	0.463	0.833	0.687	0.292	0.160	0.321	0.499	0.501	0.703	0.275	0.251	0.500	0.340	0.276	0.560	0.268	0.210	0.430
Observations	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771

Notes: The table shows estimates of dynamic treatment effects on prices paid and varieties purchased. The estimates are from regressions of the form

$$y_{igto} = \alpha + \sum_{k=1}^{3} \left(\eta_k \text{Treatment}_o^k + \kappa_k \text{Treatment}_o^k \times \text{Time}_{ito} \right) + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$$

where Treatment_o^k are dummies for office o being in treatment k; Time_{ito} is a measures of time, calendar time (scaled to be 0 at the beginning of the fiscal year and 1 at the end of the year) and/or the order of the purchase made by the office (scaled to be between 0 and 1); $\mathbf{X}_{igto}$ is a vector of controls; q_{igto} is the quantity purchased, δ_s and γ_g are strata and good fixed effects, respectively, and ε_{igto} are residuals clustered by office. Columns 1–6 estimate dynamic treatment effects on the variety purchased using the scalar measure (columns 1–3) and coarse measure (columns 4–6) described in section 5.1. Columns 7–18 estimate dynamic treatment effects on log unit prices paid, not controlling for the variety purchased (columns 7–9), or controlling for the variety purchased using the full vector of good attributes (columns 10–12), the scalar variety measure (13–15), or the coarse variety measure (16-18).

TABLE E.4: HETEROGENEITY OF TREATMENT EFFECTS BY SHARE OF BUDGET ALLOCATED TO GENERIC GOODS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Autonomy	-0.035	-0.058	-0.044	-0.051	-0.045	-0.035	-0.039	-0.056	-0.033	-0.037	-0.034	-0.048
	(0.080)	(0.064)	(0.064)	(0.068)	(0.071)	(0.049)	(0.051)	(0.057)	(0.079)	(0.055)	(0.058)	(0.064)
	[0.654]	[0.459]	[0.500]	[0.440]	[0.519]	[0.561]	[0.470]	[0.320]	[0.681]	[0.597]	[0.578]	[0.478]
Incentives	-0.004	0.009	0.011	0.029	0.007	0.015	0.018	0.029	0.002	0.015	0.017	0.033
	(0.071)	(0.052)	(0.061)	(0.061)	(0.069)	(0.050)	(0.058)	(0.060)	(0.072)	(0.052)	(0.061)	(0.063)
	[0.964]	[0.901]	[0.876]	[0.654]	[0.905]	[0.833]	[0.761]	[0.644]	[0.976]	[0.845]	[0.812]	[0.612]
Both	-0.021	-0.033	-0.027	-0.024	0.002	-0.005	-0.002	0.006	0.000	-0.009	-0.004	0.004
	(0.067)	(0.053)	(0.053)	(0.065)	(0.069)	(0.054)	(0.056)	(0.067)	(0.071)	(0.055)	(0.056)	(0.069)
	[0.762]	[0.572]	[0.626]	[0.725]	[0.980]	[0.935]	[0.976]	[0.933]	[0.993]	[0.903]	[0.934]	[0.962]
Autonomy × Generic Budget Share 14–15	-0.089	-0.057	-0.065	-0.055					-0.085	0.008	-0.035	-0.048
	(0.113)	(0.089)	(0.093)	(0.098)					(0.166)	(0.149)	(0.149)	(0.151)
	[0.405]	[0.577]	[0.477]	[0.540]					[0.625]	[0.960]	[0.838]	[0.747]
Incentives × Generic Budget Share 14–15	-0.027	-0.077	-0.069	-0.098					0.067	-0.019	0.005	-0.045
	(0.113)	(0.079)	(0.095)	(0.099)					(0.186)	(0.138)	(0.172)	(0.160)
	[0.810]	[0.465]	[0.511]	[0.313]					[0.717]	[0.910]	[0.975]	[0.790]
Both × Generic Budget Share 14–15	-0.089	-0.099	-0.083	-0.114					0.061	0.073	0.076	0.082
	(0.102)	(0.080)	(0.082)	(0.100)					(0.161)	(0.130)	(0.158)	(0.152)
	[0.348]	[0.324]	[0.322]	[0.243]					[0.764]	[0.698]	[0.707]	[0.664]
Autonomy × Generic Budget Share 15–16					-0.072	-0.099	-0.075	-0.048	-0.008	-0.104	-0.050	-0.014
					(0.102)	(0.074)	(0.076)	(0.084)	(0.151)	(0.139)	(0.133)	(0.137)
					[0.448]	[0.288]	[0.354]	[0.567]	[0.963]	[0.517]	[0.730]	[0.921]
Incentives × Generic Budget Share 15–16					-0.049	-0.089	-0.083	-0.099	-0.107	-0.070	-0.086	-0.061
					(0.110)	(0.074)	(0.090)	(0.097)	(0.182)	(0.129)	(0.164)	(0.157)
					[0.630]	[0.371]	[0.420]	[0.326]	[0.582]	[0.604]	[0.636]	[0.695]
Both × Generic Budget Share 15–16					-0.128	-0.149	-0.129	-0.167	-0.186	-0.215	-0.198	-0.244
					(0.105)	(0.083)	(0.085)	(0.103)	(0.168)	(0.138)	(0.167)	(0.158)
					[0.189]	[0.133]	[0.143]	[0.085]	[0.334]	[0.171]	[0.279]	[0.156]
Item Variety Control	None	Attribs	Scalar	Coarse	None	Attribs	Scalar	Coarse	None	Attribs	Scalar	Coarse
p(All Interactions = 0)	0.327	0.120	0.216	0.183	0.282	0.081	0.149	0.140	0.500	0.232	0.371	0.308
Observations	11,666	11,666	11,666	11,666	11,666	11,666	11,666	11,666	11,666	11,666	11,666	11,666

Notes: The table shows estimates of heterogeneous treatment effects on prices paid and by the share of the office's budget that is allocated to generic goods. The estimates are from regressions of the form

$$y_{igto} = \alpha + \sum_{k=1}^{3} \left(\eta_k \text{Treatment}_o^k + \kappa_k \text{Treatment}_o^k \times \text{BudgShare}_o \right) + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$$

where Treatment_o^k are dummies for office o being in treatment k; BudgShare_o is the share of the office's budget allocated to generic goods (in either Fiscal Year 2014–15, the first year of the experiment, or Fiscal Year 2015–16, the second year of the experiment); $\mathbf{X}_{igto}$ is a vector of controls; q_{igto} is the quantity purchased, δ_s and γ_g are strata and good fixed effects, respectively, and ε_{igto} are residuals clustered by office. Columns 1–4 estimate heterogeneity treatment effects by the office's budget share in Fiscal Year 2014–15, the first year of the experiment. Columns 5–8 estimate heterogeneity by the budget share in Fiscal Year 2015–16, the second year of the experiment. Columns 9–12 combine both years. The first column in each set does not control for item variety, the second uses all the items' attributes, the third uses the scalar variety measure, and the fourth uses the coarse variety measure.

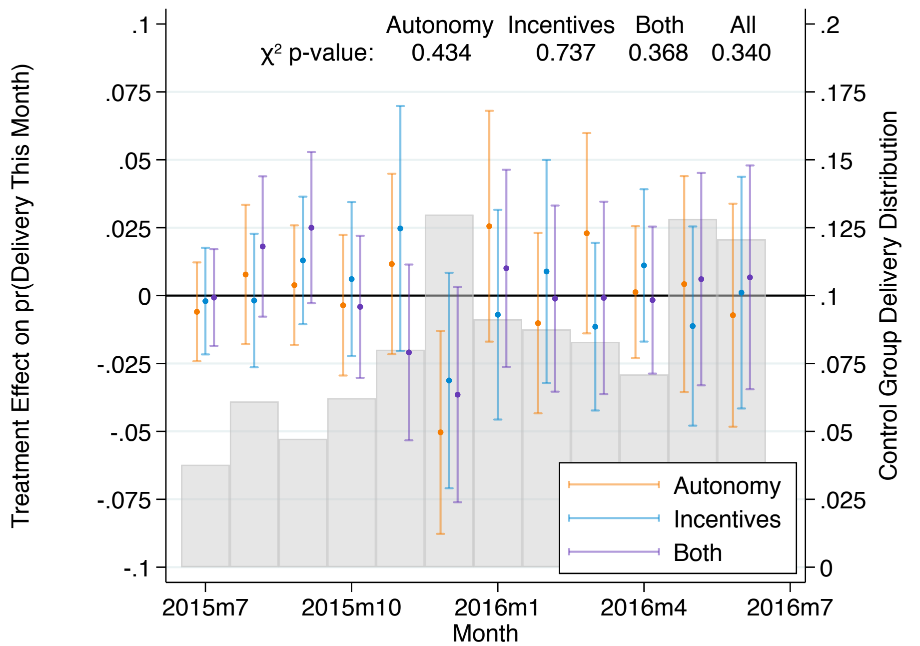
TABLE E.5: TREATMENT EFFECTS ON DEMAND FOR GOODS

Item	Treatment Effect			Joint Test
	Autonomy	Incentives	Both	All = 0
Toner	14.2 (290.62)	103.1 (294.41)	32.6 (292.44)	0.05 [0.985]
Ice Block	-6.3 (21.16)	-39.1* (21.43)	-12.3 (21.29)	1.32 [0.266]
Towel	-13.3 (12.53)	4.8 (12.70)	-15.1 (12.61)	1.25 [0.291]
Soap/Detergent	-324.0 (785.13)	11.0 (795.36)	368.4 (790.04)	0.28 [0.843]
Duster	-14.2 (11.59)	18.0 (11.74)	-16.8 (11.66)	3.91 [0.008]
Wiper	-1.3 (9.10)	22.0** (9.22)	-7.4 (9.16)	4.08 [0.007]
Lock	6.1 (20.10)	10.4 (20.36)	-17.1 (20.23)	0.75 [0.519]
Pen	54.8 (54.41)	75.8 (55.12)	19.3 (54.76)	0.78 [0.503]
Envelope	14.7 (11.18)	-4.8 (11.32)	-7.6 (11.25)	1.66 [0.172]
Printer Paper	157.1 (187.33)	254.9 (189.77)	-140.4 (188.50)	1.79 [0.147]
Register	-212.7 (357.08)	-62.3 (361.74)	68.2 (359.32)	0.24 [0.870]
Stapler	-11.8 (7.91)	-8.9 (8.01)	-13.2* (7.96)	1.09 [0.353]
Staples	-1.4 (2.87)	0.7 (2.91)	1.3 (2.89)	0.34 [0.800]
Calculator	-9.8 (8.01)	-11.5 (8.11)	-12.9 (8.06)	1.03 [0.378]
File Cover	27.7 (25.39)	-29.4 (25.72)	10.4 (25.55)	1.83 [0.139]
Stamp Pad	5.7 (4.25)	5.6 (4.30)	-1.4 (4.28)	1.58 [0.193]
Photocopying	22.5 (50.18)	55.6 (50.84)	69.8 (50.50)	0.79 [0.501]
Broom	45.1 (47.26)	84.9* (47.87)	32.8 (47.55)	1.08 [0.355]
Coal	-26.5 (58.50)	63.8 (59.26)	67.4 (58.87)	1.33 [0.263]
Newspaper	20.9 (33.64)	0.4 (34.08)	2.4 (33.86)	0.19 [0.905]
Pipe	41.6 (33.42)	90.5*** (33.85)	16.1 (33.63)	2.79 [0.039]
Light Bulb	66.6 (94.17)	-38.7 (95.40)	-2.6 (94.76)	0.45 [0.715]
Pencil	6.4 (4.36)	-0.1 (4.42)	-2.8 (4.39)	1.69 [0.167]
Floor Cleaner	-18.3 (43.58)	-3.0 (44.15)	14.7 (43.86)	0.20 [0.893]
Sign Board/Banner	123.4 (166.19)	24.1 (168.35)	32.4 (167.23)	0.22 [0.883]
Joint F-Test	0.83 [0.704]	1.23 [0.198]	0.65 [0.911]	1.08 [0.297]

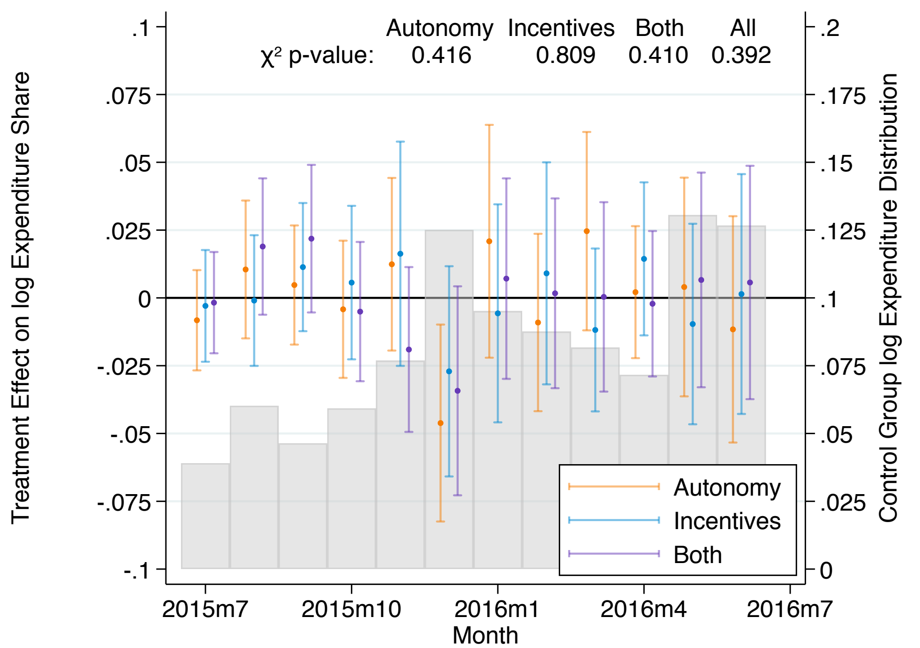
Notes: The table shows the overall treatment effects of the three treatments on the demand for different goods. We value each purchase using the counterfactual prices we estimate each purchase would have been made at had it been made by an office in the control group–the scalar variety measure. That is, for each purchase, the counterfactual expenditure is $e_{igto} = \exp(v_{igto} + q_{igto})$ where v_{igto} is the scalar good variety measure, and q_{igto} is the log number of units purchased. We then aggregate the data to the good-month-office level and estimate good-specific treatment effects by multivariate regression with the following specification for each item $e_{gto} = \sum_{k=1}^{3} \eta_{kg} \text{Treatment}_o^k + \gamma_s + \xi_t + \varepsilon_{gto}$ where e_{gto} is the quantity purchased of good g in month t by office o; the η_{kg} are good-specific treatment effects; γ_s and ξ_t are stratum and month fixed effects respectively; and ε_{gto} are residuals clustered by office. For each good, we display the estimated η_{kg} coefficients and their standard errors clustered by office, as well as the F-statistic for the hypothesis that all three η_{kg}s are equal to zero and its p-value in square brackets. In the final row, we display F-statistics for the hypothesis that each treatment has zero effect on any item, and the F-statistic on the hypothesis that none of the treatments affects any of the items.

FIGURE E.1: THE TIMING OF DELIVERIES AND EXPENDITURES IS UNAFFECTED

Panel A: Timing of Deliveries is Unaffected



Panel B: Timing of Expenditures is Unaffected



Notes: The figure shows estimates of treatment effects on the timing of deliveries and expenditure. The estimates are from seemingly unrelated regressions of the form

$$\mathbf{1}\left\{\text{Month}_i = m\right\} = \alpha + \beta_A \text{Autonomy}_i + \beta_I \text{Incentives}_i + \beta_B \text{Both}_i + \gamma_g + \gamma_s + \varepsilon_i$$

where γ_g are good fixed effects, γ_s are randomization strata fixed effects, and ε_i are residuals clustered by office. The figures show the 95% confidence intervals of the estimated β_A, β_I and β_B with p-values of χ^2 tests of the hypothesis that each treatment's effect is 0 in all months, and the hypothesis that all treatments have no effect in all months.

F Additional Results: Monitor Alignment

Definition of "Good" and "Bad" Monitors. Figure F.1 uses simple linear difference in differences specifications to explore the robustness of the position of the sharp jumps revealed by the nonparametric analysis in figure 5. Exploring one treatment at a time, we estimate treatment effect heterogeneity by interacting a dummy for facing a "bad" AG with treatment dummies $p_{igto} = \alpha + \eta \text{Treatment}_o + \zeta \text{Treatment}_o \times \text{BadAG}_o + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$ where all terms are as defined above, and we control for the scalar measure of item variety as part of $\mathbf{X}_{igto}$. Panel A of figure F.1 studies the autonomy treatment, panel B the incentives treatment, and panel C the combined treatment. The horizontal axis shows the threshold percentage of approvals in June 2015 above which an AG is considered "bad". The points show the point estimates of ζ and the bars their 95% confidence intervals using standard errors clustered by office. The gray crosses show the randomization inference p-value for the hypothesis that the effect is zero. Consistent with the non-parametric findings, the p-value falls below 0.05 at a June share of 0.22 in panels A and C, and 0.48 in panel B, so going forward we use these definitions of good/bad AGs.

Table F.1 show the robustness of our heterogeneous treatment effect estimates to our alternative ways of controlling for item variety using the same linear difference in difference specifications used in figure F.1. Panel A studies the autonomy treatment, panel B the incentives treatment, and panel C the combined treatment. As discussed in section 6.1 our proxy for the degree of misalignment of the AG is the degree to which purchase approvals are bunched at the end of the fiscal year in June 2015 (year 1 of the project). Studying one treatment at a time, we estimate treatment effect heterogeneity by interacting a dummy for facing a "bad" AG with treatment dummies $p_{igto} = \alpha + \eta \text{Treatment}_o + \zeta \text{Treatment}_o \times \text{BadAG}_o + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$. A "bad" AG is defined as a June Share above 0.22 in panels A and C, and 0.48 in panel B. Each panel is constructed in the same way, showing the η_k and ζ_k coefficients together with standard errors clustered by office in parentheses and p-values from randomization inference under the null hypothesis of no effect in square brackets. Column 1 does not control for the variety of the item being purchased. Column 2 controls for the full vector of item attributes. Column 3 uses the scalar measure of item variety. Column 4 uses the coarse measure of item variety, and column 5 uses the machine learning measure of item variety.

Alternative Proxies for Monitor Type. Table F.2 shows similar effects, particularly for the autonomy treatment, using three alternative proxies for AG type. The first alternative measure (columns 1–5) is the median weight given by respondents in the control group to the autonomy-related responses *"Only a limited number of vendors are willing to wait for delayed payment"*, *"Vendors charge higher prices for delayed payment"*, *"AG/DAO requirements are not clear and they do not clear bills without inside connections or payment of speed money"* and *"DDOs do not have enough petty cash to make purchases quickly"* when asked *"These are potential reasons for why DDOs don't achieve good value for money. In your experience how important is each of these?"* in the endline survey.

The second measure (columns 6–10) is the average of the district's monthly average delays $\sum_m \frac{1}{12}\bar{d}_m$ where $\bar{d}_m$ is the average approval delay in the district in month m. This measure avoids over-weighting delays at the end of the year (Liebman & Mahoney, 2017) when there is limited scope for monitors to cause delays. The third measure captures the speed at which delays accelerate as distance to the end of the year, and hence the scope for monitors to delay approvals, increases. For each district we regress monthly average delays on the distance to the end of the year and use the regression coefficient as our measure of how much monitors cause extra delay as the scope for them to hold up transactions increases.

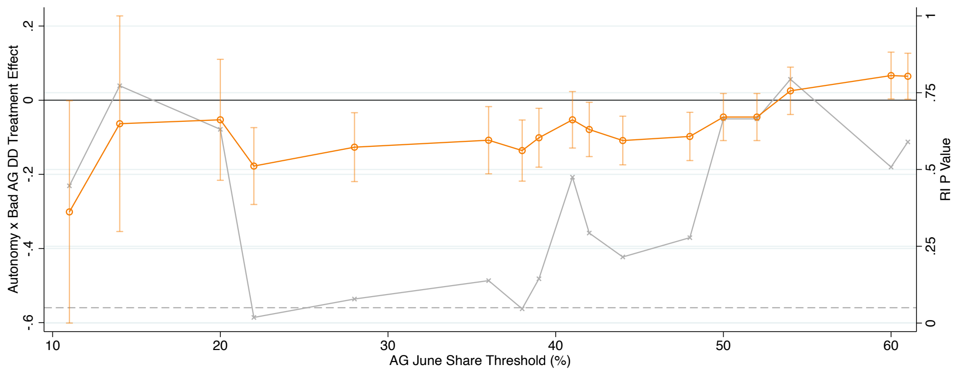
Confounders of AG Type Measures. Table F.3 shows robustness to three potential confounders our proxy for AG type may be picking up. Columns 2 & 3 control additionally for the share of submissions submitted at the end of the fiscal year in case POs submitting transactions for approval late is the real driver of the share of transactions approved late in the year. Columns 4 & 5 control for the average delay POs experience, in case the hold-up at the end of the year is driven by general delays at the AG. Columns 6 & 7 control for a measure of the PO's type in case places with bad AGs are matched with particularly good or bad POs. We estimate PO fixed effects using the year-1 data and the alternative measure is a dummy for the fixed effect being negative (below average). Note that since the incentives treatment was in place in year 1 the coefficients for the incentives and combined groups cannot be interpreted as heterogeneity by PO type. In all cases our estimates of the heterogeneity of the treatment effects by AG type are unaffected. Consistent with our findings on the overall effects in section 5.3, we find no heterogeneity of the treatment effects on the variety of items purchased or on the quantities demanded.

Non-Price Outcomes. Table F.4 shows the results of estimating the linear difference in differences specification with the scalar, coarse or machine-learning measure of item variety as outcomes, and shows no significant heterogeneity in the treatment effects. Table F.5 shows the results of estimating an extended version of equation (E.1) by multivariate regression. Specifically, for each item, we estimate $e_{gto} = \sum_{k=1}^{3} \left(\eta_k \text{Treatment}_o^k + \zeta_k \text{Treatment}_o^k \times \hat{\theta}_{AG,o} \right) +$

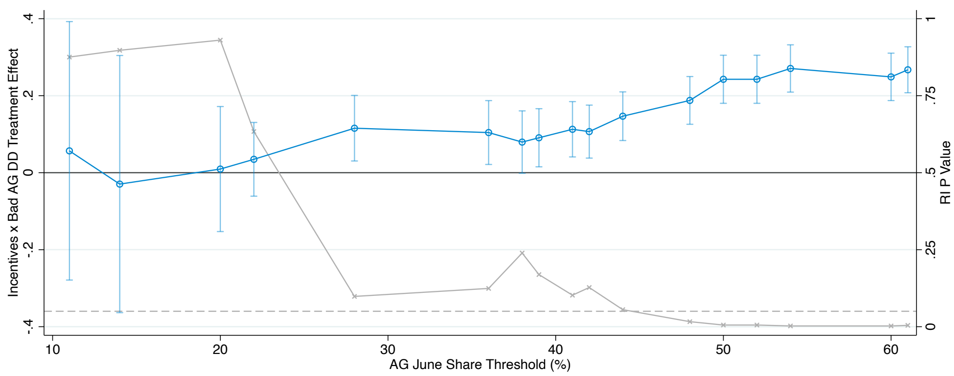
$\gamma_s + \xi_t + \varepsilon_{gto}$. We find no consistent evidence that either the linear or interaction terms imply that the treatments affected the quantity demanded, regardless of the misalignment of the AG.

FIGURE F.1: HETEROGENEITY BY AG TYPE: DIFFERENCE IN DIFFERENCES

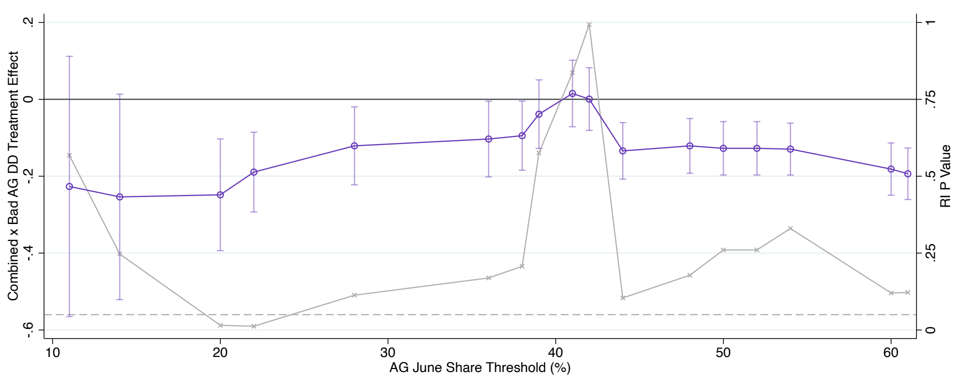
Panel A: Autonomy



Panel B: Incentives



Panel C: Combined



Notes: The figure shows how difference in difference estimates of the heterogeneity of treatment effects by monitor type change as the definition of a "bad" AG is changed. As discussed in section 6.1 our proxy for the degree of misalignment of the AG is the degree to which purchase approvals are bunched at the end of the fiscal year in June 2015 (year 1 of the project). Studying one treatment at a time, we estimate treatment effect heterogeneity by interacting a dummy for facing a "bad" AG with treatment dummies $p_{igto} = \alpha + \eta \text{Treatment}_o + \zeta \text{Treatment}_o \times \text{BadAG}_o + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$ where all terms are as defined above, and we control for the scalar measure of item variety in $\mathbf{X}_{igto}$. Panel A studies the autonomy treatment, panel B the incentives treatment, and panel C the combined treatment. Each panel is constructed in the same way. The horizontal axis shows the threshold percentage of approvals in June 2015 above which an AG is considered "bad". The points show the point estimates of ζ and the bars their 95% confidence interval using standard errors clustered by office. The gray crosses show the randomization inference p-value for the hypothesis that the effect is zero. We pick our definition of a bad AG as the threshold at which the p-value falls below 0.05: 0.22 in panels A and C, and 0.48 in panel B.

TABLE F.1: DIFFERENCE IN DIFFERENCES DESIGN

Panel A: Autonomy					
	(1)	(2)	(3)	(4)	(5)
Autonomy	0.072	0.019	0.042	0.056	0.068
	(0.072)	(0.054)	(0.053)	(0.063)	(0.072)
	[0.502]	[0.769]	[0.502]	[0.449]	[0.416]
Autonomy × Bad AG	-0.231	-0.183	-0.178	-0.210	-0.226
	(0.083)	(0.061)	(0.063)	(0.073)	(0.083)
	[0.018]	[0.016]	[0.018]	[0.013]	[0.019]
Item Variety Control	None	Attribs	Scalar	Coarse	ML
p(All = 0)	0.006	0.002	0.006	0.003	0.004
Observations	5,798	5,798	5,798	5,798	5,798
Panel B: Incentives					
	(1)	(2)	(3)	(4)	(5)
Incentives	-0.102	-0.111	-0.116	-0.094	-0.103
	(0.037)	(0.031)	(0.032)	(0.034)	(0.037)
	[0.004]	[0.004]	[0.004]	[0.014]	[0.012]
Incentives × Bad AG	0.125	0.165	0.188	0.115	0.125
	(0.076)	(0.075)	(0.064)	(0.071)	(0.076)
	[0.016]	[0.077]	[0.016]	[0.199]	[0.173]
Item Variety Control	None	Attribs	Scalar	Coarse	ML
p(All = 0)	0.006	0.011	0.006	0.075	0.072
Observations	5,413	5,413	5,413	5,413	5,413
Panel C: Combined					
	(1)	(2)	(3)	(4)	(5)
Combined	0.090	0.044	0.060	0.079	0.088
	(0.059)	(0.057)	(0.053)	(0.057)	(0.059)
	[0.336]	[0.559]	[0.336]	[0.235]	[0.194]
Combined × Bad AG	-0.240	-0.181	-0.189	-0.222	-0.239
	(0.075)	(0.069)	(0.064)	(0.071)	(0.075)
	[0.012]	[0.027]	[0.012]	[0.008]	[0.007]
Item Variety Control	None	Attribs	Scalar	Coarse	ML
p(All = 0)	0.010	0.017	0.010	0.013	0.015
Observations	5,546	5,546	5,546	5,546	5,546

Notes: The table shows heterogeneity of treatment effects by the degree of misalignment of the district's accountant general. Panel A studies the autonomy treatment, panel B the incentives treatment, and panel C the combined treatment. As discussed in section 6.1 our proxy for the degree of misalignment of the AG is the degree to which purchase approvals are bunched at the end of the fiscal year in June 2015 (year 1 of the project). Studying one treatment at a time, we estimate treatment effect heterogeneity by interacting a dummy for facing a "bad" AG with treatment dummies $p_{igto} = \alpha + \eta \text{Treatment}_o + \zeta \text{Treatment}_o \times \text{BadAG}_o + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$. A "bad" AG is defined as a June Share above 0.22 in panels A and C, and 0.48 in panel B. Each panel is constructed in the same way, showing the η_k and ζ_k coefficients together with standard errors clustered by office in parentheses and p-values from randomization inference under the null hypothesis of no effect in square brackets. Column 1 does not control for the variety of the item being purchased. Column 2 controls for the full vector of item attributes. Column 3 uses the scalar measure of item variety. Column 4 uses the coarse measure of item variety, and column 5 uses the machine learning measure of item variety.

TABLE F.2: ALTERNATIVE MEASUREMENT OF MONITOR TYPE

	Survey Responses					Weighted Delays					Delay Acceleration				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Autonomy	0.335	0.091	0.146	0.228	0.331	0.272	0.099	0.149	0.262	0.271	0.143	0.045	0.066	0.117	0.141
	(0.153)	(0.114)	(0.110)	(0.142)	(0.152)	(0.186)	(0.143)	(0.149)	(0.165)	(0.185)	(0.094)	(0.069)	(0.075)	(0.080)	(0.094)
	[0.089]	[0.507]	[0.277]	[0.186]	[0.073]	[0.244]	[0.567]	[0.389]	[0.193]	[0.231]	[0.183]	[0.586]	[0.441]	[0.213]	[0.195]
Incentives	-0.066	-0.064	-0.094	-0.036	-0.068	-0.061	-0.124	-0.151	-0.030	-0.064	-0.034	-0.061	-0.091	-0.024	-0.036
	(0.102)	(0.097)	(0.088)	(0.096)	(0.103)	(0.174)	(0.146)	(0.148)	(0.158)	(0.175)	(0.082)	(0.066)	(0.070)	(0.072)	(0.083)
	[0.631]	[0.588]	[0.378]	[0.780]	[0.601]	[0.770]	[0.478]	[0.387]	[0.871]	[0.761]	[0.725]	[0.448]	[0.257]	[0.777]	[0.718]
Combined	0.055	-0.019	-0.001	0.036	0.054	0.116	-0.044	0.002	0.117	0.119	0.192	0.039	0.113	0.160	0.193
	(0.091)	(0.091)	(0.087)	(0.102)	(0.091)	(0.191)	(0.162)	(0.168)	(0.189)	(0.189)	(0.086)	(0.075)	(0.074)	(0.083)	(0.086)
	[0.623]	[0.875]	[0.992]	[0.772]	[0.634]	[0.641]	[0.828]	[0.995]	[0.607]	[0.618]	[0.054]	[0.611]	[0.178]	[0.101]	[0.052]
Autonomy × AG Misalignment	-0.894	-0.391	-0.478	-0.655	-0.885	-0.003	-0.002	-0.002	-0.003	-0.003	-0.017	-0.010	-0.011	-0.015	-0.017
	(0.304)	(0.222)	(0.216)	(0.280)	(0.302)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.006)	(0.005)	(0.005)	(0.005)	(0.006)
	[0.023]	[0.129]	[0.067]	[0.051]	[0.022]	[0.103]	[0.297]	[0.193]	[0.074]	[0.095]	[0.014]	[0.078]	[0.064]	[0.014]	[0.015]
Incentives × AG Misalignment	0.089	0.059	0.146	0.039	0.093	0.000	0.001	0.001	0.000	0.000	0.001	0.002	0.005	-0.000	0.001
	(0.209)	(0.189)	(0.177)	(0.196)	(0.210)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)
	[0.761]	[0.781]	[0.499]	[0.884]	[0.732]	[0.836]	[0.561]	[0.442]	[0.931]	[0.819]	[0.862]	[0.655]	[0.371]	[0.996]	[0.852]
Combined × AG Misalignment	-0.291	-0.163	-0.174	-0.271	-0.291	-0.002	-0.000	-0.001	-0.002	-0.002	-0.019	-0.009	-0.013	-0.018	-0.019
	(0.192)	(0.180)	(0.172)	(0.207)	(0.192)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.006)	(0.005)	(0.005)	(0.006)	(0.006)
	[0.214]	[0.416]	[0.372]	[0.276]	[0.211]	[0.443]	[0.841]	[0.718]	[0.392]	[0.419]	[0.006]	[0.151]	[0.033]	[0.009]	[0.006]
Item Variety Control	None	Attribs	Scalar	Coarse	ML	None	Attribs	Scalar	Coarse	ML	None	Attribs	Scalar	Coarse	ML
p(All = 0)	0.072	0.067	0.040	0.083	0.043	0.145	0.111	0.114	0.073	0.129	0.001	0.015	0.007	0.004	0.002
Observations	10,172	10,172	10,172	10,172	10,172	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771	11,771

Notes: The table shows heterogeneity of treatment effects by the degree of misalignment of the district's accountant general using three alternative proxies for the monitor's type. The first alternative measure (columns 1–5) is the median weight given by respondents in the control group to the autonomy-related responses *"Only a limited number of vendors are willing to wait for delayed payment"*, *"Vendors charge higher prices for delayed payment"*, *"AG/DAO requirements are not clear and they do not clear bills without inside connections or payment of speed money"* and *"DDOs do not have enough petty cash to make purchases quickly"* when asked *"These are potential reasons for why DDOs don't achieve good value for money. In your experience how important is each of these?"* in the endline survey. The second measure (columns 6–10) is the average of the district's monthly average delays $\sum_m \frac{1}{12}\bar{d}_m$ where $\bar{d}_m$ is the average approval delay in the district in month m. This measure avoids over-weighting delays at the end of the year (Liebman & Mahoney, 2017) when there is limited scope for monitors to cause delays. The third measure captures the speed at which delays accelerate as distance to the end of the year, and hence the scope for monitors to delay approvals, increases. For each district we regress monthly average delays on the distance to the end of the year and use the regression coefficient as our measure of how much monitors cause extra delay as the scope for them to hold up transactions increases. We estimate treatment effect heterogeneity by interacting treatment dummies with the misalignment measures M_o: $p_{igto} = \alpha + \sum_{k=1}^{3}\left(\eta_k \text{Treatment}_o^k + \zeta_k \text{Treatment}_o^k \times M_o\right) + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$. The table shows the η_k and ζ_k coefficients together with standard errors clustered by office in parentheses and p-values from randomization inference under the null hypothesis of no effect in square brackets. within each set of five columns, the first column does not control for the variety of the item being purchased. Column 2 controls for the full vector of item attributes. Column 3 uses the scalar measure of item variety. Column 4 uses the coarse measure of item variety, and column 5 uses the machine learning measure of item variety.

TABLE F.3: ROBUSTNESS TO POTENTIAL CONFOUNDERS

		Late Submissions		Average Delay		Good PO	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Autonomy	0.001	-0.051	0.037	-0.110	0.125	-0.100	-0.020
	(0.057)	(0.049)	(0.074)	(0.158)	(0.182)	(0.037)	(0.061)
	[0.990]	[0.358]	[0.650]	[0.559]	[0.601]	[0.024]	[0.748]
Incentives	-0.070	-0.024	-0.060	-0.153	-0.191	0.068	0.019
	(0.036)	(0.049)	(0.049)	(0.174)	(0.165)	(0.046)	(0.053)
	[0.074]	[0.642]	[0.279]	[0.432]	[0.305]	[0.206]	[0.765]
Combined	0.037	-0.043	0.072	-0.226	-0.050	-0.039	0.032
	(0.050)	(0.056)	(0.070)	(0.221)	(0.230)	(0.043)	(0.054)
	[0.501]	[0.520]	[0.371]	[0.414]	[0.870]	[0.443]	[0.617]
Autonomy × Bad AG	-0.129		-0.138		-0.161		-0.117
	(0.065)		(0.065)		(0.064)		(0.059)
	[0.084]		[0.069]		[0.030]		[0.086]
Incentives × Bad AG	0.151		0.140		0.141		0.103
	(0.059)		(0.061)		(0.062)		(0.055)
	[0.019]		[0.035]		[0.045]		[0.094]
Combined × Bad AG	-0.155		-0.162		-0.157		-0.103
	(0.061)		(0.061)		(0.064)		(0.059)
	[0.021]		[0.018]		[0.033]		[0.118]
Autonomy × Alternative Measure		-0.115	-0.118	0.000	-0.001	0.039	0.026
		(0.145)	(0.146)	(0.002)	(0.002)	(0.060)	(0.059)
		[0.488]	[0.477]	[0.854]	[0.611]	[0.557]	[0.689]
Incentives × Alternative Measure		0.008	-0.026	0.001	0.001		
		(0.148)	(0.144)	(0.002)	(0.002)		
		[0.958]	[0.878]	[0.508]	[0.522]		
Combined × Alternative Measure		-0.113	-0.120	0.002	0.001		
		(0.178)	(0.175)	(0.002)	(0.002)		
		[0.593]	[0.549]	[0.594]	[0.760]		
p(All = 0)	0.002	0.229	0.009	0.335	0.007	0.012	0.002
Observations	11,771	11,771	11,771	11,771	11,771	11,771	11,771

Notes: The table shows robustness of the heterogeneity of treatment effects by the degree of misalignment of the district's accountant general to the inclusion of alternative explanations. As discussed in section 6.1 our proxy for the degree of misalignment of the AG is the degree to which purchase approvals are bunched at the end of the fiscal year in June 2015 (year 1 of the project). We estimate treatment effect heterogeneity by interacting treatment dummies with a dummy for facing a "bad" AG and with alternative explanation measures $p_{igto} = \alpha + \sum_{k=1}^{3} \left(\eta_k \text{Treatment}_o^k + \zeta_k \text{Treatment}_o^k \times \text{BadAG}_o + \xi_k \text{Treatment}_o^k \times \text{AlternativeMeasure}_o\right) + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$. A "bad" AG is defined as a June Share above 0.22 for the autonomy and combined treatments, and 0.48 for the incentive treatment. We show the η_k ζ_k and ξ_k coefficients together with standard errors clustered by office in parentheses and p-values from randomization inference under the null hypothesis of no effect in square brackets. Column 1 does not include any alternative hypothesis. Columns 2 and 3 consider heterogeneity caused by transactions being *submitted* for approval late. The alternative measure is the share of transactions submitted for approval at the end of the year (May and June). Columns 4 and 5 consider heterogeneity caused by general delays in monitors approving purchases. The alternative measure is the average delay between submission and approval. Columns 6 and 7 consider heterogeneity caused by the effectiveness of the procurement officers (POs) rather than the monitors. We estimate PO fixed effects using the year-1 data and the alternative measure is a dummy for the fixed effect being negative (below average). Note that since the incentives treatment was in place in year 1 the coefficients for the incentives and combined groups cannot be interpreted as heterogeneity by PO type.

TABLE F.4: HETEROGENEITY OF EFFECTS ON ITEM VARIETY BY MONITOR TYPE

	(1)	(2)	(3)
Autonomy	0.042	0.017	0.013
	(0.037)	(0.039)	(0.013)
	[0.298]	[0.666]	[0.362]
Incentives	0.022	0.021	0.008
	(0.022)	(0.026)	(0.011)
	[0.367]	[0.488]	[0.467]
Combined	0.047	0.111	0.004
	(0.028)	(0.038)	(0.012)
	[0.134]	[0.005]	[0.732]
Autonomy × Bad AG	-0.075	-0.013	-0.021
	(0.042)	(0.043)	(0.016)
	[0.093]	[0.772]	[0.244]
Incentives × Bad AG	-0.048	0.011	-0.010
	(0.038)	(0.040)	(0.018)
	[0.277]	[0.785]	[0.601]
Combined × Bad AG	-0.067	-0.073	0.001
	(0.037)	(0.043)	(0.016)
	[0.095]	[0.104]	[0.948]
Variety Measure	Scalar	Coarse	ML
p(All = 0)	0.503	0.130	0.936
Observations	11,771	11,771	11,771

Notes: The table shows heterogeneity of the treatment effects on the variety of the items purchased by the degree of misalignment of the district's accountant general. As discussed in section 6.1 our proxy for the degree of misalignment of the AG is the degree to which purchase approvals are bunched at the end of the fiscal year in June 2015 (year 1 of the project). We interact our treatment dummies with dummies for facing a "bad" AG in the following specification: $v_{igto} = \alpha + \sum_{k=1}^{3} \left(\eta_k \text{Treatment}_o^k + \zeta_k \text{Treatment}_o^k \times \text{BadAG}_o \right) + \mathbf{X}_{igto}\beta + \rho_g q_{igto} + \delta_s + \gamma_g + \varepsilon_{igto}$. A "bad" AG is defined as a June Share above 0.22 for the autonomy and combined treatments, and 0.48 in the incentives treatment. We showing the η_k and ζ_k coefficients together with standard errors clustered by office in parentheses and p-values from randomization inference under the null hypothesis of no effect in square brackets. The dependent variable in Column 1 is the scalar measure of item variety. In column 2 we study the coarse measure of item variety, and in column 3 we study the machine-learning (ML) measure of item variety.

TABLE F.5: HETEROGENEITY OF EFFECTS ON DEMAND BY MONITOR TYPE

Item	Linear Term			Bad AG Interaction			Linear	Interactions
	Autonomy	Incentives	Both	Autonomy	Incentives	Both	All = 0	All = 0
Toner	119.1	252.5	-800.8	-140.6	-495.0	1172.9**	1.73	2.43
	(489.98)	(335.64)	(490.34)	(559.01)	(528.17)	(559.61)	[0.158]	[0.063]
Ice Block	-24.5	-52.6**	-12.1	26.0	44.5	0.3	1.61	0.56
	(35.68)	(24.44)	(35.71)	(40.71)	(38.46)	(40.75)	[0.184]	[0.643]
Towel	-17.0	-3.0	11.0	5.1	25.8	-36.6	0.52	1.65
	(21.14)	(14.48)	(21.15)	(24.11)	(22.78)	(24.14)	[0.669]	[0.176]
Soap/Detergent	143.4	-37.7	-59.4	-651.4	142.4	597.1	0.01	0.19
	(1324.31)	(907.16)	(1325.28)	(1510.88)	(1427.52)	(1512.52)	[0.999]	[0.906]
Duster	-15.3	13.1	-21.0	1.8	16.0	6.0	1.27	0.20
	(19.55)	(13.39)	(19.56)	(22.30)	(21.07)	(22.33)	[0.284]	[0.896]
Wiper	10.4	19.6*	-11.7	-16.2	7.6	5.9	2.04	0.58
	(15.35)	(10.52)	(15.36)	(17.52)	(16.55)	(17.53)	[0.106]	[0.631]
Lock	49.6	-16.3	-7.1	-60.7	86.0**	-14.3	1.38	3.36
	(33.88)	(23.21)	(33.91)	(38.65)	(36.52)	(38.70)	[0.246]	[0.018]
Pen	34.8	64.9	-30.0	29.2	35.7	70.3	0.57	0.17
	(91.78)	(62.87)	(91.85)	(104.71)	(98.94)	(104.83)	[0.637]	[0.919]
Envelope	30.7	-12.9	-52.6***	-21.8	25.8	63.5***	5.92	4.96
	(18.84)	(12.90)	(18.85)	(21.49)	(20.30)	(21.51)	[0.000]	[0.002]
Printer Paper	477.2	-30.0	-544.3*	-437.2	917.4***	570.9	2.99	4.57
	(315.71)	(216.26)	(315.94)	(360.18)	(340.31)	(360.57)	[0.030]	[0.003]
Register	-10.6	-55.1	-293.1	-280.0	-31.6	506.8	0.09	0.37
	(602.28)	(412.57)	(602.72)	(687.13)	(649.22)	(687.88)	[0.964]	[0.778]
Stapler	-0.3	-13.5	6.8	-16.3	14.8	-28.4*	1.14	2.02
	(13.34)	(9.14)	(13.35)	(15.22)	(14.38)	(15.23)	[0.330]	[0.109]
Staples	3.4	-1.6	-2.0	-6.6	7.1	4.6	0.47	1.82
	(4.84)	(3.31)	(4.84)	(5.52)	(5.21)	(5.52)	[0.702]	[0.142]
Calculator	-12.9	-19.0**	-15.3	4.6	24.5*	3.6	1.48	0.95
	(13.51)	(9.25)	(13.52)	(15.41)	(14.56)	(15.42)	[0.219]	[0.416]
File Cover	12.1	-8.3	17.1	21.2	-68.4	-9.7	0.14	0.94
	(42.83)	(29.34)	(42.86)	(48.86)	(46.16)	(48.91)	[0.936]	[0.422]
Stamp Pad	5.2	6.0	-11.4	0.9	-1.4	14.2*	2.48	1.20
	(7.17)	(4.91)	(7.17)	(8.17)	(7.72)	(8.18)	[0.059]	[0.309]
Photocopying	-98.8	66.5	144.2*	169.5*	-31.5	-103.2	2.78	2.25
	(84.61)	(57.96)	(84.67)	(96.53)	(91.20)	(96.63)	[0.040]	[0.080]
Broom	54.2	92.9*	0.4	-12.6	-26.5	45.5	1.14	0.17
	(79.71)	(54.60)	(79.77)	(90.94)	(85.92)	(91.04)	[0.331]	[0.915]
Coal	70.5	63.9	9.0	-135.9	-3.6	80.8	0.42	1.02
	(98.66)	(67.58)	(98.73)	(112.56)	(106.35)	(112.68)	[0.742]	[0.385]
Newspaper	63.7	-0.8	-13.3	-60.1	2.3	21.4	0.64	0.47
	(56.75)	(38.87)	(56.79)	(64.74)	(61.17)	(64.81)	[0.590]	[0.703]
Pipe	206.5***	135.5***	47.6	-234.1***	-151.9**	-48.5	7.05	5.76
	(56.31)	(38.57)	(56.35)	(64.24)	(60.70)	(64.31)	[0.000]	[0.001]
Light Bulb	206.9	-103.1	-269.0*	-192.1	204.5	375.3**	2.85	3.11
	(158.75)	(108.74)	(158.87)	(181.11)	(171.12)	(181.31)	[0.036]	[0.025]
Pencil	3.1	-0.3	-0.8	4.7	0.8	-2.7	0.10	0.21
	(7.35)	(5.04)	(7.36)	(8.39)	(7.93)	(8.40)	[0.962]	[0.886]
Floor Cleaner	-44.4	-11.1	-70.7	38.2	26.9	121.4	0.34	0.70
	(73.51)	(50.35)	(73.56)	(83.86)	(79.24)	(83.95)	[0.799]	[0.552]
Sign Board/Banner	352.4	-10.7	-114.0	-319.7	105.7	203.4	0.90	0.79
	(280.28)	(192.00)	(280.49)	(319.77)	(302.13)	(320.12)	[0.440]	[0.502]
Joint F-Test	1.19	1.52	1.04	1.16	1.35	1.45	1.42	1.50
	[0.236]	[0.047]	[0.403]	[0.261]	[0.116]	[0.068]	[0.010]	[0.003]

Notes: The table shows the results of estimating an extended version of equation (E.1) by multivariate regression. Specifically, for each item, we estimate $e_{gto} = \sum_{k=1}^{3} \left(\eta_k \text{Treatment}_o^k + \zeta_k \text{Treatment}_o^k \times \hat{\omega}_s \right) + \gamma_s + \xi_t + \varepsilon_{gto}$ on data aggregated up to the office $\times$ month $\times$ good level. To aggregate the data, we weight each purchase by our scalar measure of item type, which can be interpreted as the price we predict the item would cost had it been bought in the control group in year 1. For each purchase, demand is $e_{igto} = \exp\left(q_{igto} + h_{igto}\right)$, where q_{igto} is the log number of units purchased in purchase i, and h_{igto} is the scalar item type measure, and we sum over all purchases of good g in month t by office o to create e_{gto}.