NBER WORKING PAPER SERIES

PAST PERFORMANCE AND PROCUREMENT OUTCOMES

Francesco Decarolis
Giancarlo Spagnolo
Riccardo Pacini

Working Paper 22814
http://www.nber.org/papers/w22814

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2016

We are grateful to Nick Bloom, Luis Cabral, Liran Einav, Rosa Ferrer, Stefano Gagliarducci, Ricard Gil, Luigi Guiso, Hugo Hopenhayn, Neale Mahoney, Enrico Moretti, Gustavo Piga, Piero Sabbatini, Stephane Saussier, Steve Tadelis, Paola Valbonesi and to the participants at the seminars at Barcelona GSE, Berkeley, FGV Rio de Janeiro, Stanford University, Stockholm University, University of Bologna and University of Paris I for their useful suggestions. We are thankful to the firm running the experiment and to the Italian Procurement Authority for providing us with the data. Decarolis thankfully acknowledges financial support from the European Research Council (ERC Starting Grant 679217). Spagnolo is grateful to the Wallander and Hedelius foundation for financial support (Handelsbanken P2013-0162). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Past Performance and Procurement Outcomes

Francesco Decarolis, Riccardo Pacini and Giancarlo Spagnolo[*]

May 18, 2020

Abstract

Reputational incentives may be a powerful mechanism for improving supplier performance and limiting the perverse effect of price competition on contract execution. We analyze a unique experiment run by a large utility company in Italy which introduced a new vendor rating system scoring its suppliers' past performance and linking it to the award of future contracts. We study responses in both price and performance to the announcement of the switch from price-only to price-and-rating auctions. Average performance improves from 25 percent to 90 percent of the audited parameters. Improvements involve all parameters and suppliers, are long-lasting (for at least 10 years after the initial experiment) and are reflected in higher service quality by the utility. Contract prices do not significantly change overall, but we find evidence of lower prices right after the announcement when suppliers compete to win contracts to get scored, and of higher prices, once they have established a good reputation. We then argue that supplier moral hazard is the main force behind our findings. The main takeaway from this study is that the gains from curtailing supplier moral hazard may be higher than those from always bolstering price competition, and that a reputational mechanism based on objective past performance can be a powerful tool to achieve this goal.

JEL: H57, D47, K12

Keywords: Public procurement, past performance, reputation, vendor rating

[*] Decarolis: Bocconi University, *francesco.decarolis@unibocconi.it.* Pacini: Agenzia del Demanio, *riccardo.pacini01@agenziademanio.it.* Spagnolo: SITE-Stockholm School of Economics, EIEF and Tor Vergata, *spagnologianca@gmail.com.* We are grateful to Nick Bloom, Luis Cabral, Liran Einav, Rosa Ferrer, Stefano Gagliarducci, Ricard Gil, Luigi Guiso, Hugo Hopenhayn, Neale Mahoney, Rob Metcalfe, Enrico Moretti, Gustavo Piga, Piero Sabbatini, Stephane Saussier, Steve Tadelis, Paola Valbonesi and to the participants at the seminars at Barcelona GSE, Berkeley, FGV, Stanford University, Stockholm University, University of Bologna and University of Paris I for their useful suggestions. We thank Acea for sharing its data. We also thankfully acknowledge financial support from both the European Research Council (ERC Starting Grant 679217, Decarolis) and the Wallander and Hedelius foundation (Handelsbanken P2013-0162, Spagnolo).

I Introduction

Reputational forces linking future business to past performance are a pillar of large sectors of the economy, from business-to-business negotiations to transactions over electronic platforms.[1] In this paper, we exploit a unique setting of a large scale firm experiment to document for the first time, and in considerable detail, the effects of reputational incentives on performance and prices in a centralized, auctions-based public procurement market.

In private procurement, past performance indicators have always affected the selection of suppliers and their behavior because private buyers are free to act upon them by refraining from selecting suppliers with a poor track record. In public procurement, however, this type of discretional management practice is typically limited: the need to prevent corruption led lawmakers around the world to ensure that open auctions, where bidders receive equal treatment, are used as often as possible, even if supplier track records differ considerably. Unfortunately, as is well known, competitive auctions can be a problematic mechanism in the context of contract procurement: with incomplete contracts, bolstering competition at the bidding stage might come at the cost of poor ex post performance.[2] Balancing this price versus performance trade-off, and – specifically – how the use of past performance can contribute to that, is a fundamental, yet unsolved, problem of public procurement.

Our study contributes to the analysis of this problem by exploiting a very rich set of data related to the introduction of a past performance monitoring system in the procurement practices of a large Italian multi-utility company, Acea. This company provides water and power to a vast area in central Italy that includes its capital, Rome. In 2007, Acea started an "experiment" with a new vendor rating system to understand how to improve contractual performance (in terms of quality and safety) in the execution of the construction jobs it

[1] See Tadelis (2016) for a recent survey of reputational mechanisms in electronic platforms. On business-to-business negotiations involving the sale of complex goods, see Banerjee and Duflo (2000).

[2] A classic reference is Spulber (1990) which shows that in the construction sector, where contracting is typically imperfect, open competition spurs adverse selection and ex post opportunism of contractors. More generally, the limits of competitive auctions in this type of settings have been shown theoretically by – among others – Manelli and Vincent (1995), Zheng (2001), Bajari and Tadelis (2001) and Burguet, Ganuza and Hauk (2012) and confirmed empirically by Bajari, McMillan and Tadelis (2009), Decarolis (2014), Liebman and Mahoney (2016), Lewis-Faupel et al. (2016) and Kang and Miller (2019).

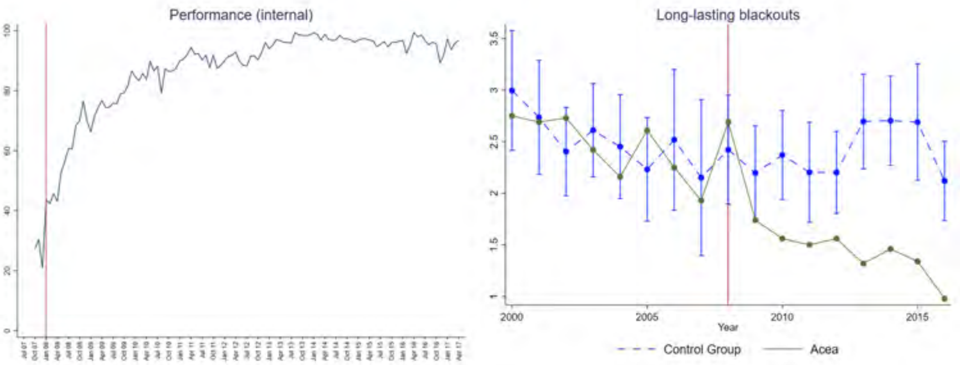
awards for the maintenance and upgrade of the electricity grid.[3] Acea's engineering unit laid down a list of 136 observable parameters measuring both quality and safety features of the job carried out by its suppliers. Acea's auditors, who used to perform worksite inspections and then write paper-based memos, were given tablets embedded with a software to record the scores on these parameters. Only a few months later, Acea made its first public statement explaining to its suppliers that the results of the new audit system would be converted into a numerical "reputation index" and that such an index would be used to award new contracts after a few more months of data collection.

This timing allows us to study the evolution of both price and performance around the time when the new audit system was publicly *announced*. As extensively discussed below, focusing on the period when the new price-and-reputation auction system was only announced, but not yet implemented, has a series of advantages concerning the interpretation of the empirical analysis. Moreover, it is right after the announcement period that the most interesting dynamic involving both price and performance takes place. Hence, in the first part of the paper, we analyze how compliance with the monitored parameters monitored evolved in response to the timing of the five public announcements. Using audit data for the years 2007 to 2009, we find clear evidence of a substantial change in contractors' behavior: compliance in the 136 parameters increased from 25 percent before the first announcement to more than 80 percent before the first auction took place under the new price-and-performance award criterion. We find that essentially all active suppliers improved their compliance in similar ways and they did so strategically, with compliance increasing relatively more for those parameters with higher weights in the computation of the reputation index. The striking increase in compliance for the audited parameters is clearly shown by the left panel of Figure

[3] For Acea, this was an experiment in the sense that it introduced the new audit system only for a subset of contracts (all of those involving public illumination and electrical-substation works, but none of those related, for instance, to water delivery) and its stated goal was to learn whether the new audit system could be beneficial for its overall procurement. While this experiment does not satisfy all the characteristics of an ideal field experiment (List and Reiley, 2008), it is nevertheless a very useful natural experiment that, to the best of our knowledge, allows us to do the first quantitative analysis of the effects of a reputational mechanism in public procurement. For the distinction among types of firm-level experiments see Bloom et al. (2014) who state that "in personnel economics there is a tradition of exploiting changes in firm policies initiated by a CEO (a natural experiment such as Lazear (2000))." In our case, the experiment was autonomously decided by Acea's CEO, but we had an active role in its design and implementation, in the collection and analysis of the first few years of data, as well as in the five communications to the suppliers discussed below.

1: full compliance for all parameters, by all the firms audited in a given month would set the blue line equal to 100, but we see that the compliance level in 2007 is only between 20 percent and 30 percent. The vertical, red line marks the date of the first announcement: it is evident how performance improves after this date. Moreover, the long time series of available data provides a rare opportunity to observe the long lasting impact of this reform which entails average compliance settling at around 90 percent. As discussed at the end of this section, this is the case even after a legal controversy led to the dismissal of the price-and-reputation system and its replacement with a hybrid mechanism.

Figure 1: Internal and External Performance Measures



Note: the left panel shows the monthly average (weighted) compliance recorded in Acea's audits. A value of 100 would thus imply that, across all the audits taking place in that month, all suppliers were compliant on all parameters inspected. The right panel displays the yearly evolution of one of the external performance measures: the number of long-lasting blackouts (i.e., those lasting 3 hours or more) per client, for both Acea (in green) and other utilities (in blue). The red line indicates the date of Acea's first announcement.

In the right panel of Figure 1, we report the evolution of the average number of blackouts per client for both Acea and the other electricity distributors in Italy. This is one of the external performance measures to which we have access. As discussed below, these are all measures that are not part of the scored parameters, but that represent socially-relevant outcomes associated with the quality of the distribution service. Improvements in the performance of Acea's suppliers should lead to improvements in the external measures, although possibly with a time lag. This is indeed what we observe in the right panel: the number of blackouts experienced by Acea's customers declines (green line), both in absolute terms and relative to those of the clients of the other utility company in the control group (blue line).

Improvements in blackouts are necessarily more gradual than improvements in the internal performance measure due to technological features. Even if the quality of the maintenance work on the grid suddenly improves, observing greater reliability in the electricity grid requires a large portion of it to undergo work executed according to the higher quality standards. Moreover, a second feature linked to supplier behaviour that we document below contributes to explaining this slower improvement in the external performance measures: supplier improvement is not uniform across parameters. It is faster for parameters carrying more weight in the reputation index (which are mostly related to worksite safety) and those that are cheaper to improve quickly. A similar pattern characterizes all the external measures involving the electricity sector. On the contrary, no improvements are found for the Acea's water distribution service, which was not part of the experiment. Both the total amount of water leakage experienced by Acea grows over time and its relative performance to the other water distributors does not improve.

Given this evidence on improved performance, in the second part of the analysis we ask what its cost was and whether it was worth it. We find that the cost was quite limited and most likely worth the benefits in terms of extra quality and safety. This part of the analysis takes advantage of a second dataset containing information on all the contracts awarded by both Acea and all other Italian public buyers, mostly utilities, procuring the same type of contracts. We use the variation across procurers and over time to develop a difference-in-differences identification strategy. We find that, if we consider the date of the first announcement to be the one characterizing the occurrence of the policy change, then there is no significant effect on the price paid by Acea. More specifically, by looking at any symmetric window of time around the first announcement, prices remain stable on average. However, when we extend the empirical model to account for the evolution of compliance, the price response appears more nuanced. Using the results from the first part of the analysis, we partition the period after the first announcement into a first phase when compliance grows, and a second phase when it flattens out at high levels. When we extend the baseline difference-in-differences model to account for different behaviors in these two phases, we see that the original finding of no effect results from the combined effects of prices declining when compliance improves, but prices increasing after compliance stabilizes.

We interpret this evidence as suggestive of a first phase in which suppliers compete harder to win contracts: since only contract winners can be audited, winning is required to earn (or improve) the reputation index. Winning a contract has thus the additional benefit of improving the chances of winning future contracts. After all contractors have earned a high reputation index, however, this benefit is outweighed by the increased cost of high compliance, and auction prices become correspondingly higher. The estimates indicate that these effects cancel out each other.

The lack of any significant cost increase coupled with evidence of improved performance allow us to assert that the reform was a cost effective improvement over the previous status quo. We formalize this argument by using our estimates to quantify the savings produced by reducing both the probability of deadly accidents and the duration of blackouts. By using the OECD figures for the value of a statistical life together with the same statistical model employed by Acea's engineers to map the relationship between changes in parameter compliance and the occurrence of fatal accidents, we estimate that the benefit from increased compliance on the safety parameters ranges between €3.5 and €5.3 million per year. Furthermore, from the official statistics of the electricity regulator, we associate a cost to every hour of blackouts for both residential customers and business customers: the reduction in blackouts implies a benefit of €6.6 million, 39 percent of which accrues to business customers.

The final part of the analysis studies the mechanisms driving our findings. In particular, it focuses on whether the observed effects are the result of changes in the selection of contractors which are bidding, or in their behavior, i.e., moral hazard. The evidence is definitely compatible with the latter, implying the presence of moral hazard: suppliers that are observed bidding both before and after the new rating system is announced stop offering suspiciously low prices. These are precisely the abnormal, low ball bids often associated with poor ex post contractual performance. On the other hand, we find only limited effects of selection based on three features in the data. First, while several suppliers leave the market, the timing of their exit is not associated with the announcements. Second, both the firms that leave the market and those that remain have similar bidding patterns. Third, for many observable characteristics, the firms leaving Acea's auctions are no different from the suppliers leaving the auctions of another large multi-utility company that did not participate in the experiment and that we use as a benchmark. Thus, the main result from this study is that the gains from curtailing suppliers moral hazard may be higher than those from always bolstering price competition, and that a reputational mechanism based on objective past performance can be a powerful tool to achieve this goal.

Overall, the results in this study contribute to multiple strands in the literature. At the highest level, it is related to two strands of the law and economics literature on agency problems. The first strand concerns ex ante regulation vs. ex post incentives. Shavell (1984), and the research line following from it, modeled the theoretical question of whether ex ante or ex post interventions are more effective tools for dealing with a firm engaging in potentially risky behaviours and having private information about the extent of potential hazards.[4] Acea, with its dominant position as the largest buyer in the market, is akin to a regulator that decides to bolster the role of ex post incentives to curb risky behaviors by its regulated subjects (i.e., Acea' suppliers).[5] The second strand is that of the efficiency-corruption trade-off in delegation within an organization, see Banfield (1975) as the classic reference. Price-only auctions represent rigid mechanisms where delegation to the agents (i.e., Acea's engineers) of the awarding and monitoring the contracts is minimal. The introduction of a reputation system requires delegating more powers to the auditors, thus risking that they will exploit it for personal gain. In our case, we find that increasing delegation is beneficial. Interestingly, a wave of recent papers on public procurement – see Coviello, Guglielmo and Spagnolo (2017), Carril (2019) and Decarolis et al. (2020) – all reached the same conclusions despite looking at different countries and different types of discretion, thus suggesting that public procurement regulations tend to involve too little delegation.[6]

Our findings are also related to a recent wave of studies highlighting the importance of

[4] A large body of subsequent studies have extended this original result and explored applications ranging from environmental protection to banking regulation. See, among others, Kolstad, Ulen and Johnson (1990), Rose-Ackerman (1991) and Hiriart, Martimort and Pouyet (2004).

[5] Importantly, the success of this strategy likely hinges on fact that the enforcement of ex ante (i.e., contract) clauses through penalties is limited by the well know inefficacy of the Italian civil court system. See Djankov et al. (2008) for a cross-country study and Giacomelli and Menon (2016) for Italy.

[6] As discussed below, Acea's approach involved not only fostering delegation but also containing corruption risks through a mechanism of rotation and random drawing on the pool of auditor scored suppliers.

adopting a dynamic framework to understand and redesign public procurement markets.[7] Our study is the first to empirically document the strength of this approach. A remarkable feature of our results is the sheer size of the performance improvements driven by the market design intervention. This runs against the typical "revenue equivalence's curse" by which the strategic behavior of bidders undoes most (if not all) of the benefits that the designer intended to achieve with its intervention. Crucial for our result is the repeated nature of the procurement process. In this respect, this study is close in spirit to those of Jofre-Bonet and Pesendorfer (2003), Marion (2017) andChassang and Ortner (2016) which stress that a dynamic approach to repeated procurement is key to understand the role supplier behavior.

Another strand of the literature to which our paper contributes is that on the design and use of contract audit measures. The detailed performance measures, from random audits by centrally managed inspectors we have access to, relate our paper to the work on public procurement by Olken (2007) on Indonesia and Colonnelli and Prem (2017) on Brazil. Related to the issue of corruption, Bandiera, Prat and Valletti (2009) show that, in the context of Italian procurement, corruption concerns could be less of a priority than inefficient procurement. Our paper is a step in the same direction, as it identifies and quantifies the costs, benefits, and channels of an improvement in the efficient organization of a public procurement system.

Finally, on the policy side, this study contributes to the hotly debated issue of the supplier past performance in public procurement. We will return to this debate in the conclusions, but we shall anticipate that controversies over the proper use of past performance have been rampant in the European Union. Indeed, due to growing concerns within Acea's legal counsel that the experiment described below, and especially the use of price-and-reputation auctions was in breach of the EU procurement directives, the experiment was ultimately abandoned after just 36 price-and-reputation auctions were held in the space of about one year. Nevertheless, the new audit system was never abandoned and is still used today. The auctions system returned to being price-only, but was combined with a new provision

[7] A few theoretical studies have argued in favor of the positive role that past performance and reputation may play in improving contract performance in repeated public procurement under imperfect contracting. See, among others, Kim (1998), Doni (2006) and Calzolari and Spagnolo (2009).

allowing Acea's inspectors to block the payments to the supplier if the audits on a worksite revealed major violations. This latter feature is important in understanding the persistence of the observed performance improvements which indeed come not just from the changes in physical capital and managerial behaviour (similar to the cases reviewed in Brandon et al. (2017)), but also from changes in the environment within which suppliers operate.

II The Experiment

The context of the experiment is that of a multi-utility company, Acea s.p.a., offering electricity and water services to about 1.6 million customers, both private households and business establishments, in the Rome area. The firm is vertically integrated, owning and operating the majority of its generation, transmission and distribution systems. From this point of view, it is very similar to some of the largest US power operators such as the Los Angeles Department of Water and Power (LADWP), ComEd (Chicago), BGE (Baltimore) and PECO (Philadelphia).[8] As shown in Table 1, all of these firms spend significant resources every year on works aimed at preserving the operational efficiency of its power grid.

Table 1: Comparison with U.S. Multi-Utility Providers

	ACEA	LADWP	ComEd	BGE	PECO
Total Employees (000)	5.0	9.4	6.8	3.3	2.6
Power Customers (mln)	1.6	1.4	3.8	1.3	1.6
Power Grid (000/miles)	19	14	90	26	14
Total Turnover (bln/$)	3.2 (2.1)	4.4 (3.3)	4.9	3.1	3.0
Power Supply (TWh)	11	26	86*	29*	36*
Works on Power Grid Works (mln/$)	206	318	2,400	500	475

*Note: Acea and LADWP figures on employees and turnover include the water business too. BGE and PECO figures on employees and turnover include the gas business too. All values are for 2015. Values with a * symbol are estimates: the supply is estimated proportionally to the customers out of the total supply of all Exelon subsidiaries (195TWh). For the total Turnover (bln $), the values in parenthesis refer to power only.*

In 2015, Acea spent about US $200 million on procuring the kind of works which are the

[8] The external validity of what can be learned from a firm-level experiment is a typical concern in the literature (Bloom et al. (2014)). In our case, it is thus reassuring to observe that Acea is similar to both some other major operators active in the US, such as the multi-utility companies of the four US cities mentioned above, and to the other companies providing the same services in Italy, as discussed below.

focus of the experiment in this study. The jobs typically entail the maintenance, upgrade and replacement of transformers, poles, underground cables, underground vaults, station transformers, distribution and receiving stations.[9] These are all works exposing workers to safety hazards linked to electricity-induced accidents. In 2007, after these risks materialized in some deadly accidents, the company's management decided to take action to enhance the safety (and quality) standards in contract execution.

The main goal was to introduce an objective measure of contractual performance (combining elements of both safety and quality), with the plan of using its ratings in the award stage of future procurement processes. That is, two processes that had been separated until then – contract procurement and ex post auditing – were going to become interdependent.[10] At the same time, it was also decided to leave out from this experiment the water sector in order to have a benchmark against which to evaluate the effectiveness of the new system. Within the electricity sector, two groups of jobs – works related to public illumination and electricity distribution – were considered sufficiently homogeneous to define a list of items to assess contractual performance. A total of 136 parameters were identified for this goal. Table 2 reports how these 136 parameters are divided into 12 categories, further divided into 2 macro classes: "safety" (51 parameters; 7 categories) and "quality" (85 parameters; 5 categories). For instance, "Equipment and machinery," the first category in Table 2, comprises 5 parameters involving the adequacy of both the formal documentation and the physical condition of equipment and machinery.[11] Parameters in this category are quite general and can be inspected for essentially all work sites. Other categories, instead, involve parameters specific to a subset of contracts only. For instance, the 25 parameters in "Underground works" involve exclusively jobs on underground wires and electrical substations.

[9] Moreover, specific investment is required to integrate increasing amounts of intermittent renewable generation resources and transformational technologies such as energy storage, electric vehicles, and other aspects of the smart grid. Hence, resource planning and infrastructure asset management need to be aligned to ensure ageing assets are replaced with infrastructure that is able to meet new system requirements and maintain reliability with a modern generation mix.

[10] Contract auditing was conducted even before this experiment begun, but purely for legal reasons linked to contract enforcement and not to determine the awarding of subsequent contracts. Although the outcomes of these audits could have been used to enforce penalties, penalties were rarely enforced in this market, partly to avoid legal disputes and partly not to disrupt cooperation within the buyer-seller relationship, crucial for this type of work according to Acea's management.

[11] While important for the safety of the work site, these features might also influence the quality of the

Table 2: Reputation Index Components

Class	Category	Parameters	
		Number	Avg. Weight
Safety			
	Equipment and machinery	5	8.4
	Documentation	9	6.9
	Works execution	8	8.8
	Personnel	4	9.3
	Works site regularity	10	8.2
	Works site safety	10	9.4
	H.T. works site controls	5	8.8
Quality			
	Works on joints	19	5.7
	Customer relationship mgnt	3	7.3
	Air works	25	6.7
	Underground works	25	6.0
	Works on transformer station	13	6.2

Note: The table reports the two classes and 12 categories in which the 136 parameters are subdivided. For each category, the first number reported is the number of parameters in that category, while the second is the average RI weight across these parameters.

The system works as follows. Scores are collected by teams of rotating auditors (Acea's engineers) in one or more visits to the work sites, with a score assigned to each of the 136 parameters.[12] The score is 1 if the value is "compliant," zero if "not compliant" or "n/a" if it is impossible to inspect. In an average audit, 34 parameters are scored with either a zero or a 1. The scores on the individual parameters are then aggregated into a unique reputation index (RI). Each parameter is associated with a weight, ranging from 2 to 10. The RI is calculated as a weighted average across a predefined time span:

$$RI = \frac{\sum_{i=1}^{m} \sum_{j=1}^{136} p_{ij} u_j}{\sum_{j=1}^{136} u_j}, \tag{1}$$

with $p_{ij} \in \{0, 1\}$ indicating the score in each of the $j \in \{1, .., 136\}$ parameters, with $u_j \in \{2, 3, ..., 10\}$ being the weight attached to parameter j and m being the set of audits considered (at each point in time, these are the audits in the previous 12 months). Hence,

work executed, thus making the distinction between the two classes of quality and safety blurry.

[12] Two system features entail a randomization: which contracts are audited and which engineers from Acea are assigned to the team inspecting each work site are both determined through a process of random drawing. Thus, a single contract might be audited one or more times and by the same or different engineers.

RI ranges from 0 to 1 and entails no differential discounting of the m audits.[13]

We will refer to the elements composing the RI as "internal" performance measures. All other measures of quality and safety which are monitored but not included in the RI calculation, such as the number or duration of blackouts and the number or intensity of workers' accidents, will be indicated as "external" performance measures. The availability of the latter will be particularly useful to assess the extent to which improvements in the internal measures might be the result of the shifting of efforts to those parameters in the RI, possibly leading to performance worsening along non-monitored dimensions.

On October 16, 2007, the Acea's engineers conducted their first audit with the new auditing system. For the following 3 months audits were conducted in this way and, to the firms receiving these inspections, it was explained that Acea had simply decided to modernize its auditing process: the former paper-based memos where inspectors described the state of the worksite had been abandoned in favor of a digitalized recording system based on a set of 136 parameters. Indeed, Acea's engineers used a tablet pc to record and transmit the scores recorded during their worksite visits. The true motivation for the switch to the digitalized audits was later revealed to the suppliers in a public meeting held on December 20, 2007 ($t1$). On this occasion, Acea announced to its contractors the intention to switch its contract procurement system from price-only auctions to price-plus-performance auctions. In the latter, the winner would be the firm with the highest score S calculated as:

$$S = w_{price}(1 - \frac{\text{Price offered}}{\text{Reserve price}}) + (1 - w_{price})RI, \tag{2}$$

where w_{price} is the weight assigned to price relative to that assigned to the RI. Hence, from a status quo of a $w_{price} = 1$, the new system would entail switching to a $w_{price} = 0.75$. Equation (2) is a form of linear scoring rule auction that gives incentives to perform well in

[13] Similar systems exist in other utility companies. For instance, the LADWP *Contractor Performance Program* states that: *"A Contractor Performance Scorecard system will be maintained on all contractors that have been identified by end users and Contract Administrators of having not met the terms of the contract/purchase order. An "infraction point" will be assessed against the contractor when contractual terms are not met and are communicated to the contractor. Each point will stay on a contractors record in Supply Chain Services for a 12 month period after infraction. If a particular contractor receives a total of 3 or more points within a 12 month period, that contractor may be debarred from bidding with the Department of Water & Power for a period of up to 5 years."*

contract execution to accumulate RI points valuable for future auctions. Both in this first meeting with its suppliers and in 4 follow up meetings held in the following 13 months, Acea extensively explained this new system, showed simulation of how a firm would benefit from higher RI and updated each firm by (privately) informing it of its current RI, as well as (publicly) disclosing the distribution of RI across all suppliers.

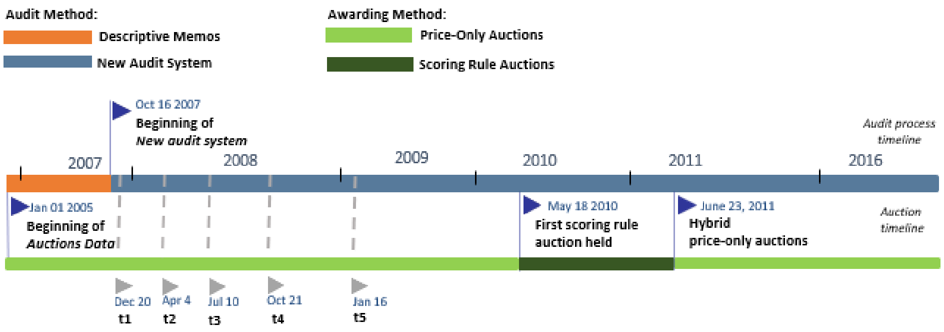
The final two crucial and interrelated aspects worth discussing involve new entrants and legal constraints. Regarding the former, Acea announced that the RI, calculated as in equation (1), would apply exclusively to those bidders audited at least 7 times in the previous 12 months.[14] Otherwise, a bidder would be assigned a RI equal to the average RI of the bidders in the auction. The same averaging rule was going to be used for new entrants (i.e., firms never audited). Regarding the legal constraints, they likely represented the most significant concern for most of the Acea personnel involved in this experiment. Without entering into the complexities of this issue, we shall remark that while Italian, and European Union, regulations encourage the use of scoring rule auctions (which are known as MEAT, most economically advantageous tenders), the parameters in the scoring formula must pertain to the bids and not the bidders. The basic logic is that allowing supplier-based scores would create a risk of favoritism, which would be detrimental to the establishment of a single European public procurement market. To what extent a system like that in equation (2) can be reconciled with the laws is, however, an open and intensely debated issue in the community of scholars and practitioners involved in EU procurement.[15]

The differing views on this topic explain the peculiar timing of the implementation of equation (2). After just 36 scoring rule auctions in the space of about a year, the new management team, that in the meantime had replaced the one under which the experiment had been designed, decided to abandon the combined auctions-audits system, officially ter-

[14] This requirement concerns the number of audits and not the number of contracts, as a supplier can be audited multiple times for the same contract.

[15] Directive 18/2004, art. 2 required that "contracting authorities shall treat economic operators equally and non-discriminatorily and shall act transparently." Under art. 54 of the Directive 17/2004, reputation indicators can be used if based on measurable parameters that are verifiable by third parties and agreed upon by contractors. The EU Court of Justice, however, ruled that contracting authorities, when evaluating quality with MEAT, should only consider the object of the tender and not the bidder's characteristics (like past performance), see judgments in cases C-488/01 and C-31/87.

Figure 2: Timeline



Note: Timeline of the changes in the auditing (top bar) and auctioning (bottom bar) systems. Acea's five announcements of the future switch to equation (2) are marked with t1,...,t5, with t1 being the first announcement date, and t2,...,t5 the dates of follow up meetings where Acea provided additional explanation to its suppliers regarding this new system.

minating the experiment. As illustrated in Figure 2, the auditing system that was changed in October 16, 2007 was maintained and is still in use (see upper timeline). Instead, the auction system was first switched from price-only to a scoring rule on May 18, 2010, but then returned to price-only auctions in June 2011 (see lower timeline). We refer to this latter system as *hybrid price-only auction*. In fact, to avoid worsening the contractual performance, Acea's inspectors were given new powers to block the contract execution if major violations were detected during the audits. To resume the job, the supplier would need to give proof that all violations were being fixed.

The timing described in Figure 2 has important implications for our analysis. It implies that we shall consider the period leading up to the switch to the scoring rule in May 2010 as a period when suppliers credibly expected the introduction of price-plus-performance auctions. In this period, they competed to win contracts under price-only auctions but were already building their stock of RI. Clearly, in this period the RI could not act as a barrier to entry since bids were just price discounts, but it could have already affected the firms' decisions of whether to participate in an auction and which price to offer. Hence, for this period from December 2007 up until May 2010 we can cleanly study the effects produced by the *announcement* of a new, reputation-based system. The subsequent periods are also interesting to discuss, but we remark that their interpretation is less straightforward for two main reasons. First, it is less credible that all firms correctly anticipated the timing with which the scoring rule auctions would be removed. Second, as we shall see below, both the switch to scoring rule auctions and to the hybrid system took place after suppliers had already modified in a substantial way those features that were driving low quality and safety before the experiment started.

III Data

The analysis is based on three sets of data. The first comes from Acea and contains audit data covering the internal performance measures recorded through the new auditing system. The second combines data from Acea and Telemat, a large provider of public tender data, and contains auction data, covering bidding and other auction-related information. The third comes from the public authorities supervising the power and water sectors and contains the external performance measures.

A. Acea's Audits: Internal Performance Measures. The first dataset contains all of Acea's audits under the new system, from its introduction in October 2007 until April 2017. There are 302,634 scores assigned to each parameter inspected during 8,974 audits involving 634 contracts and 73 different contractors. Recall that, since the subset of worksites inspected in each given week is randomly drawn at the beginning of that week, a contract might receive no inspections at all or multiple inspections during its life. Although the shorter-lasting contracts might be rarely observed in the data, the level of detail of this dataset offers a rare opportunity to evaluate how contractual performance evolved over a 10-year period. Table 3 offers some initial descriptive evidence by reporting summary statistics, aggregating parameters at the level of the 12 categories. The table shows that there is substantial heterogeneity in the frequency with which different parameters are scored: very few contracts entail features that allow inspectors to check parameters in the "Customer relationship mgnt" category whilst, at the opposite end of the spectrum, parameters in the "Works site regularity" and "Works site safety" categories are systematically assessed.

The table also reports the average share of compliant parameters (i.e., those scored with

Table 3: Summary Statistics for the Acea's Audits (Internal Performance Measures)

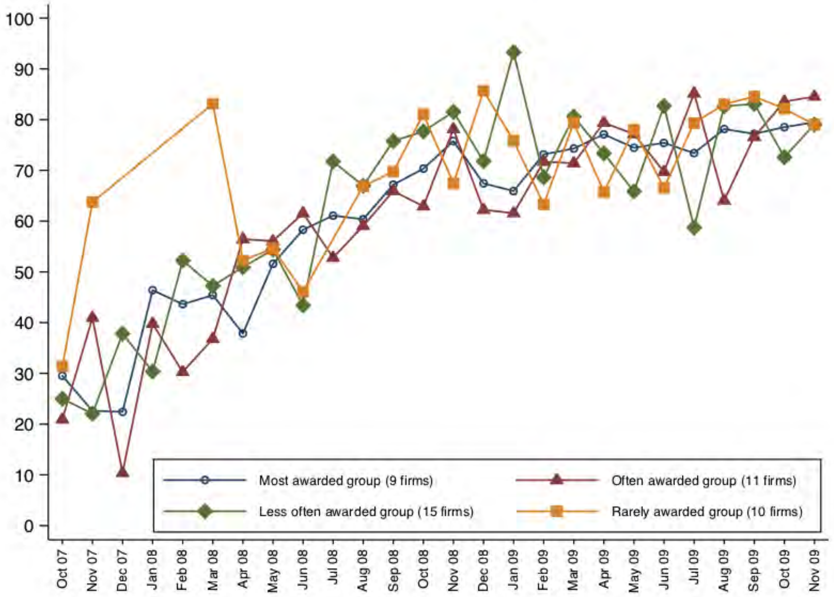
Class	Category	Share Compliant Parameters				Number of
		Pre t1	Post t1	SR period	Post SR	observations
Safety						
	Documentation	0.33	0.65	0.84	0.93	53,121
	Equipment and machinery	0.70	0.93	0.96	0.95	44,266
	H.T. works site controls	.	0.79	0.93	0.97	2,507
	Personnel	0.32	0.67	0.91	0.96	21,513
	Works execution	0.19	0.84	0.97	0.98	30,663
	Works site regularity	0.10	0.61	0.84	0.94	59,531
	Works site safety	0.31	0.75	0.92	0.96	78,338
Quality						
	Works on joints	1	0.96	1	1	1,746
	Customer relationship mgnt	1	0.94	.	1	85
	Air works	.	0.98	1	1	146
	Underground works	0.40	0.69	0.91	0.89	10,450
	Works on transformer station	1	1	1	1	268

Note: The 136 parameters audited are partitioned into the 12 categories and 2 classes indicated in the first two columns. For each of the four subperiods in which the sample is split, the share of compliant parameters indicates the share of scores equal to 1, over the sum of all scores that are either zero or 1.

a 1 over all those scored with either zero or 1). The share is reported separately for each of the 12 categories and for four time periods: before the suppliers were informed of the true motivation for the digitalized audits (Pre $t1$), after they received this information but before the introduction of scoring auctions (Post $t1$), during the scoring auctions (SR Period) and after the hybrid price-only auctions (Post SR). Across nearly all categories, there is a sharp increase between the Pre $t1$ and Post $t1$ periods. The increase is more moderate in the latter periods. For instance, for the two most audited classes "Works site regularity" and "Works site safety", the increase between the first two periods is stunning: from 10 percent to 61 percent and from 31 percent to 75 percent respectively. By contrast, the change observed between the latter periods is more modest: from 84 percent to 94 percent and from 92 percent to 96 percent respectively. Indeed, this decreasing rate at which performance improves was already observed in Figure 1 and will be further analyzed in the next section.

Finally, another important angle of the analysis, to which we will return in the discussion of the mechanisms presented at the end of the paper, regards the role of moral hazard versus adverse selection among Acea suppliers. Figure 3 offers an insight into what will be

Figure 3: Evolution of Contractors' Performance over Time



Note: the four lines show the progress of the average compliance to the parameters of the Reputation Index, calculated on a monthly basis, for 4 different groups of firms. Each observation is thus the average of all the scores obtained by all of the firms in a given firm group and a given month. The groups are formed on the basis of the firm's success in concluding contracts. The line with circle markers represents the "most awarded" firms, triangles are for the "often awarded" group, diamonds are for "the less awarded group" and squares are for "the rarely awarded group."

discussed more extensively there: all suppliers improved their performance, albeit with a different timing. By pooling suppliers into 4 groups depending on the frequency with which they win, the positive trend in compliance is evident for all of them. The higher performance by those suppliers winning less often should not be surprising: these are the firms bidding less aggressively, thus winning less, but delivering higher quality.

B. Auctions Data. The second dataset contains data on the awarding of public procurement auctions. By combining internal Acea data with data from a private provider of data on public contracts (Telemat), we obtained a dataset covering the universe of auctions held between 2005 and 2016 for the type of maintenance jobs involved in Acea's experiment.[16]

[16] These jobs belong to a well-defined contract category identified by the Italian regulation as "$OG10$," which makes it feasible to select comparable projects across different buyers. Furthermore, by using textual search methods, we were able to separate $OG10$ contracts into those involving public illumination and those involving electrical substations. Finally, to ensure contract comparability, we trimmed a few particularly large or small contacts (i.e., all of those with a reserve price below €10,000 or above €2.5 million).

The data include the object of the contract, the reserve price, the award price and date, the identity of both the procurer and the winning contractor, and various other information on the call for tenders, such as the award procedure and criterion. For a subset of auctions, we integrate the data with the information on losing bids and on the subsequent life of the contracts using data from the authority supervising public contracts (ANAC).

Table 4: Summary Statistics for the Auctions Data

Panel (a): Pre-announcements (01/2005-11/2007)

	Acea			Control		
	Mean	SD	N	Mean	SD	N
Winning Discount	21.73	10.51	172	21.30	10.19	2020
Winning Bid	516.1	428.6	172	445.5	522.4	2020
Length (days)	401.6	179.1	172	327.8	340.4	1788
Num. Bids	10.69	4.305	172	-	-	-
Public Illumination	0.180	0.386	172	0.266	0.442	2020
Central Region	1	0	172	0.202	0.402	2020
Municipal Firm	1	0	172	0.390	0.488	2020

Panel (b): Post-announcements & before SR period (12/2007-03/2010)

	Acea			Control		
	Mean	SD	N	Mean	SD	N
Winning Discount	18.99	10.40	138	22.95	11.60	2247
Winning Bid	516.1	313.5	138	384.9	468.1	2247
Length (days)	385.9	146.7	138	354.1	1106.8	1741
Num. Bids	11.21	4.337	138	-	-	-
Public Illumination	0.232	0.424	138	0.265	0.442	2247
Central Region	1	0	138	0.197	0.398	2247
Municipal Firm	1	0	138	0.395	0.489	2247

Panel (c): SR and hybrid price-only periods (04/2010-12/2016)

	SR period			Post SR		
	Mean	SD	N	Mean	SD	N
Winning Discount	28.76	7.292	35	28.16	6.631	159
Winning Bid	513.2	260.5	35	884.6	616.1	159
Length (days)	421.3	98.24	35	421.3	183.6	159
Num. Bids	13.52	2.336	35	12.42	4.568	159
Public Illumination	0.629	0.490	35	0.245	0.432	159

Note: selected summary statistics for the auction data. "Control" sample consists of auctions held by CAs other than Acea. Panel (a) covers auctions held before t1 by both Acea and Control units; Panel (b) covers auctions held at or after t1 (and before the switch to SR) by both Acea and Control units; Panel (c) covers Acea's auctions held under either the SR (left panel) or the hybrid price-only (right panel) systems. The definition of the variables is as follows: Winning Discount is the discount (over the reserve price) offered by the winning supplier, Winning Bid is the price bid by the winning supplier, Length is the contractual duration of the contract in days (a contractual duration of 1 year corresponds to 250 working days), Num. Bids is the number of bids submitted, Public Illumination is a dummy equal to 1 if the contract type is classified by Acea as public illumination and zero if it is classified as work on electrical substations, Central Region is a dummy equal to 1 if the CA is located in one of Italy's Center regions and zero otherwise and Municipal Firm is a dummy equal to 1 if the CA is a multi-utility company that is (at least partially) owned by the municipality in which it operates. The last two variables are not reported for Panel (c) as they are both always equal to 1 for the Acea's auctions.

Table 4 reports summary statistics for the auction data, dividing them in three panels. The top panel describes the data during the pre-announcement period (i.e., 01/2005-11/2007). The middle panel covers the data after the first announcement ($t1$), but before the SR was implemented. The bottom panel presents statistics for the later periods, after the SR was introduced. The first two panels report the data for both Acea and the control group, the last one reports data for Acea only, but separately for the SR and post-SR periods. The main outcome variable for the price analysis below is the winning discount.[17] The comparison of the top and bottom panels of Table 4 indicates that the average winning discount in Acea's auctions declines, from 21.73 percent to 18.99 percent, while it grows in the Control group's auctions, from 21.30 percent to 22.95 percent. This suggests that the prices paid by Acea might have increased after the first announcement. The validity of the control group is clearly illustrated by the common trend observed in top-left panel of Figure 4 for the period before the first announcement ($t1$ is marked in the figure by the red, vertical line). The figure also reveals a more nuanced pattern for the winning discounts after $t1$ relative to what is visible from the statistics in Panel (b): discounts first increase and then sharply decrease (soon after $t5$). The very different behavior in the control group suggests that this is likely due to Acea's reform and not to changes in market conditions. The following analysis will establish these effects formally.

Regarding the other variables reported in Table 4, there are no major differences between the top two panels, neither for Acea nor for the Control group. This is the case, for instance, for contract duration or the share of public illumination contracts.[18]

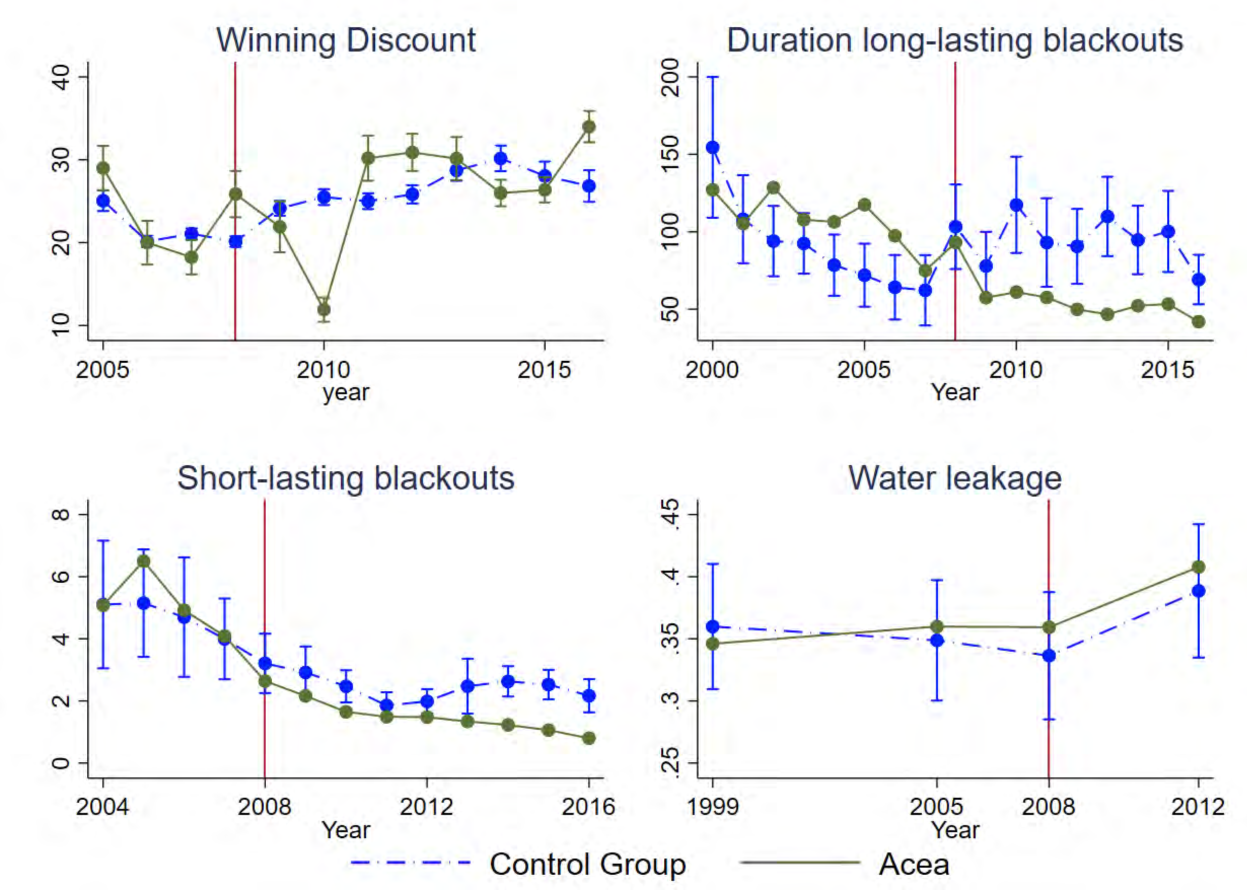
Finally, Panel (c) reports statistics for the period from the introduction of the SR onward. The Difference-in-differences strategy presented next focuses exclusively on the sample periods of Panel (a) and (b). The statistics in Panel (c) are nevertheless interesting to get a

[17] Bids are percentage discount relative to the reserve price publicized in the call for tenders. This reserve price is unlikely to be affected by Acea's reform because public buyers are not in full control of it: it is obtained by multiplying input quantities (estimated by the procurer's engineers) by their prices and summing up these products. Crucially, input prices are not the current market prices, but the list prices set every year by the region where Acea operates and used exclusively by contracting authorities to calculate reserve prices.

[18] It is important to stress that the main effort to ensure the comparability of the auctions was at the data collection stage, where we selected only auctions that, in terms of their object, were a close fit to the public illumination and electricity distribution contracts auctioned off by Acea.

sense of the longer run impacts of the reform. In particular, we observe that during the 35 auctions using the SR procedure, there is a sharp increase in the discounts relative to the earlier period and that this higher discount level is preserved during the following hybrid price-only system. As shown in Figure 4, this increase takes place in the Control group too and, hence, likely reflects some broader trend in the market. Finally, notice that the reserve price is higher in the post-SR period relative to those in the SR period: this is part of a trend in Acea's contracting in order to concentrate its demands into fewer, larger lots.

Figure 4: Evolution of Discounts and External Performance Measures



*Note: The figure illustrates external performance measures on water and electricity for both Acea (in green) and other providers (in blue). Top-left: blackout duration; top-right: the number of short-lasting blackouts; bottom-left: number of programmed power cuts; bottom-right: water leakage. In all graphs, the red, vertical line indicates the t*1 *announcement date.*

C. Regulatory Reports: External Performance Measures. In Italy, electricity and water are both partially-regulated sectors. For electricity, although only power transmission is still under a regulatory regime, the regulator (ARERA) collects detailed information on the whole sector. From ARERA we were thus able to obtain various firm-level performance measures. These yearly data range from year 2000 to 2016 and cover all low-voltage power distributors, including Acea. Herein, the main indicators of firm performance are constituted by the number and duration of blackouts and programmed power cuts. The top six rows of Table 5 report summary statistics for these external performance measures, none of which is part of the RI parameters. As discussed when presenting Figure 1 in the introduction, the external performance measures allow comparison of Acea's performance to that of other similar firms. In that figure, we plot the evolution of the number of long lasting blackouts. In Figure 4, we do the same for three other external performance measures: the blackouts duration (in minutes) and the number of blackouts lasting less than 3 hours. The observed pattern is qualitatively similar to that discussed earlier: after $t1$, Acea's performance gradually improves in both absolute and relative terms.[19] The reasons why improvements in electric grid performance occur more slowly than those in internal performance are mostly due to technological constraints: even if suppliers use higher quality joints and materials (some of the quality parameters, see Table 3), only when a large enough portion of the grid is affected will blackouts fall. In the next section, we will explore some additional features linked to Acea's suppliers' behavior that contribute to explaining the slow improvement in external performance measures.

For the water sector, we do not have Acea's internal performance measures as this sector never introduced a digitalized system like that described above. Nevertheless, external performance measures have been obtained from the environmental census of the Italian Statistical Institute (Istat). This census is performed in collaboration with the water distributors and includes information on water inflow and outflow in the distribution channel for each Italian county from 1999 to 2012. A performance measure is thus the extent of water leakage, calculated as the percentage incidence of leakage over water inflow. Although the data is released at county level, it is easy to aggregate counties in such a way as to pin

[19] In the appendix, Figure A.1 reports the analogous plots for the programmed power cuts. More programmed power cuts typically imply improved service quality as they are associated with work on the grid and they substitute unplanned blackouts. Although the plots in Figure A.1 show very small changes in Acea's planned power cuts, the estimates in the next section will suggest improvements along these measures.

Table 5: Summary Statistics for the Regulators' Reports (External Performance Measures)

VARIABLES	(1) Mean	(2) St. Dev	(3) Median	(4) Min	(5) Max	(6) N	(7) Source
Long-lasting blackouts (*num*/*LVlines*)	2.43	2.50	1.76	0	24	1,433	ARERA
Blackouts duration (*min*/*LVlines*)	94	134	49.40	0	960	1,419	ARERA
Short-lasting blackouts (*num*/*LVlines*)	2.70	3.90	1.84	0	62	1,286	ARERA
Programmed power cuts (*num*/*LVlines*)	0.6	1.24	0.30	0	29.50	1,431	ARERA
Duration programmed power cuts (*min*/*LVlines*)	65.60	114	31.20	0	989	1,428	ARERA
Low voltage users (*thousands*)	365	815	6.42	0	4,664	1,642	ARERA
Water Leakage (%)	0.33	0.09	0.32	0.15	0.74	257	ISTAT
Water users (*thousands*)	893	1,054	491	119	4,341	257	ISTAT

Note: Long-lasting blackouts and Blackouts duration are, respectively, the average number and the average duration (in minutes) of long-lasting blackouts per user, Short-lasting blackouts is the average number of short-lasting blackouts per user, Programmed power cuts and Duration programmed power cuts are, respectively, the average number and average duration (in minutes) of programmed power cuts to the low voltage grid per user, Low voltage users is the total number of low voltage grid customers (in thousands), WaterLeakage is the percentage incidence of water leakage over water inflow (Water Leakage= (Inflow-Outflow)/Inflow), while Water users is the total number of customers (in thousands).

down the water leakage level experienced by Acea. In fact, by law each county can have no more than one water distributor, so we simply aggregated up the water leakage data for all the counties served by Acea.[20] The bottom rows in Table 5 report summary statistics for the water sector, while the bottom panels of Figure 4 plots the dynamic of the water leakage indicator, separately for ACEA and other firms. There is no visual evidence of lower leakages for Acea, both in absolute terms and relative to other providers.

IV Empirical Analysis

The descriptive evidence so far shows that Acea's reform improved contract performance over the following 10 years. A careful empirical analysis is nevertheless needed to answer three questions crucial to deriving more general implications from this experiment. First, what triggered the performance improvement and, in particular, was it driven by a response to the

[20] This aggregation is performed by weighting the leakage in each of the counties served by a provider by its share of water customers relative to the total population of water customers served by the provider. County data are aggregated to mirror the "catchment areas" over which there is, by law, only one water provider.

announcement of the scoring auction? Second, what was the effect on prices of the changes in performance? Third, was the improvement in performance confined to the internal measures or did it also affect the external performance measures? These are interrelated but distinct questions that we will address through different combinations of the data described earlier and with different empirical strategies.

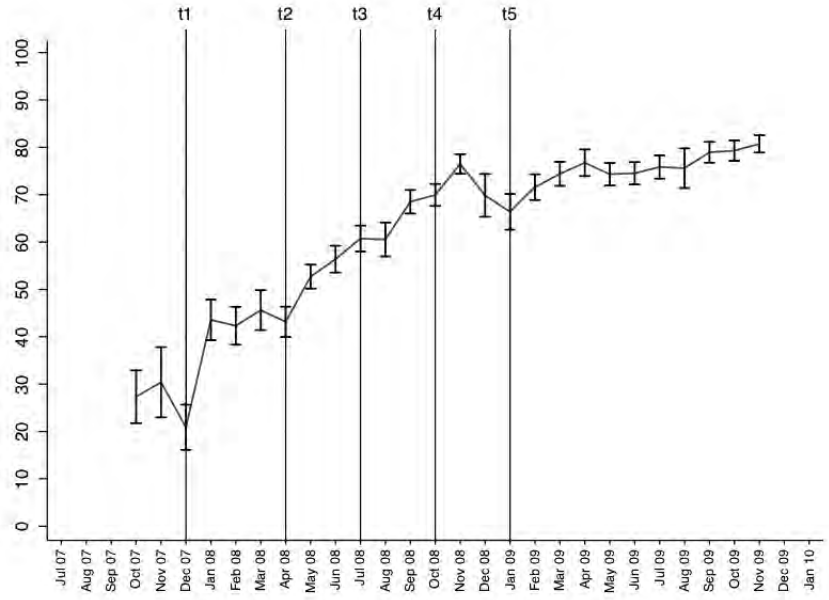
In particular, for the first question we need to rely on Acea's audit data and on their time series analysis. In fact, no comparison group is available in this case to serve as a benchmark. We will instead exploit the very clear timing of the events in which Acea presented its new procurement system to the suppliers to evaluate changes in their contract performance around these announcements. It is a different case for the following empirical questions whose answers are based on the regulatory and auctions datasets in which both Acea and similar providers are observed. We will thus follow a differences-in-differences estimation strategy:

$$O_{ft} = a_f + b_t + cX_{ft} + \beta D^{Acea*Post} + \epsilon_{ft}, \tag{3}$$

where O_{ft} is an outcome measure observed for unit f in year t. In the regulatory data, f will indicate firms, while it will refer to contracts in the auctions data. On the right hand side of the equation, a_f and b_t are fixed effects for firms and years, while X_{ft} is a matrix of controls and, finally, $D^{Acea*Post}$ is a dummy for Acea's auctions held after some pre-specified date marking the beginning of a treatment (i.e., after $t1$). The coefficient of interest is β, which thus captures the difference in external performance between Acea and other firms, after the treatment date. But what should be the treatment date(s)? The earlier description of the experiment clarify that multiple candidate dates exist. A benefit of the time series analysis from which we now move is that it provides the answer to this question, thus representing a key pillar of the following difference-in-differences estimates.

A. What Caused Performance Improvements? As discussed earlier, most of the performance improvements observed over the long run took place during the first years (see Figure 1). In Figure 5, we focus on this earlier period, zooming into the dynamics of performance after the new audit system was introduced but before the switch to scoring auctions. We also add to Figure 5 vertical bars marking each one of the Acea's announcements, $t1, ..., t5$. We can visually observe how performance jumps upward after each announcement – except $t3$ – and how its growing dynamic reduces its speed soon after $t5$. Moreover, the variance declines over time, as shown by the 95 percent confidence interval for the monthly mean.

Figure 5: Average Compliance



Note: The graph shows the monthly average compliance with the internal parameters (audits data). The average is calculated across all the scores recorded in all the audits taking place in the month of reference, weighting each parameter by its weight in the RI. The vertical lines identify each announcement date.

This graphical evidence illustrates what clearly emerged during the experiment: suppliers began improving their compliance with the audited parameters even before the scoring rule was introduced and Acea's announcements had a key role in driving this behavior. To formally show the connection between performance changes and announcement timing, Table 6 reports the results of Bai-Perron tests for the presence of structural breaks in the time series of the compliance measure in the same time window as Figure 5. In the first two columns, the variable of interest is the monthly weighted average compliance across all parameters. The next two columns restrict the parameters to those in the quality class, while the last two columns use the subset of parameters in the safety class. We report Bai-Perron tests in which we do not specify the dates of the breaks but let the test determine them, either without specifying how many breaks there are (odd numbered columns) or specifying that there are 5 breaks at unknown dates (even numbered columns). The test results are a clear indication that $t1$ is a breakpoint. As regards the other break dates, all tests allowing for an unspecified number of breaks identify a break near $t5 + 1$ (i.e., 1 month after $t5$).[21] This is also quite revealing since, by the fifth meeting, suppliers had found out that average compliance had reached a fairly high level across all active suppliers and parameters. As discussed below, this likely changed the strategic environment in the auctions, through a change in the perceived value of further improvement in compliance.

Table 6: Breakpoints in the Internal Performance Measures

	Weighted Compliance		*Quality*		*Safety*	
	F-stat breaks	5 unknown breaks	F-stat breaks	5 unknown breaks	F-stat breaks	5 unknown breaks
Number of breaks	4	5	2	5	4	5
Dates of the brakes:						
Date 1	t1	t1	t1	t1	t1	t1
Date 2	t2	t2	t3+2	t3+2	t2	t2
Date 3	t3+1	t3+1	-	t4+2	t3+1	t3+1
Date 4	t5+1	t5+1	-	t5+2	t5+7	t5
Date 5	-	t5+7	-	t5+5	-	t5+7

Note: The table reports the results of Bai-Perron tests. The variable is the monthly weighted average compliance, measured on all audited parameters (first two columns) or on the subset of quality parameters (next two columns) or safety parameters (latter two columns). We indicate as $ty + x$ a breakpoint taking place x months after Acea's announcement date ty, where $y = 1, ..., 5$. The test criterion used is that of sequential F-statistic determined breaks. Results are identical with the significant F-statistic largest breaks criterion.

In Table 7, we complement the time series evidence with estimates of linear regressions of the average monthly compliance by contract and supplier on dummy variables for the four break dates detected by the Bai-Perron test (see column (1), panel (b) of Table 6) and other controls (the share of safety parameters among those audited and whether the contract is for public illumination). Column (1) confirms the significance of all four break dates. However, more interestingly, when we gradually augment the set of regressors to include fixed effects for suppliers, contracts and months, we find that the dummy for $t1$ preserves its statistical significance and large magnitude, thus confirming its saliency. In the appendix, we also present an extensive set of robustness checks showing that the observed

[21] Either exactly at $t5 + 1$ in the case of the overall compliance, and at $t5 + 2$ for the quality parameters or at $t5$ for the safety parameters.

Table 7: Acea's Announcements and Supplier Compliance

	(1)	(2)	(3)	(4)
t1	0.200***	0.195***	0.172***	0.456**
	(0.044)	(0.042)	(0.049)	(0.200)
t2	0.065**	0.060**	0.058**	0.082
	(0.031)	(0.030)	(0.029)	(0.132)
t3+1	0.122***	0.115***	0.138***	-0.107
	(0.024)	(0.023)	(0.023)	(0.109)
t5+1	0.082***	0.067***	0.055***	0.013
	(0.018)	(0.017)	(0.020)	(0.068)
Firm Fixed Effects	No	Yes	Yes	Yes
Contract Fixed Effects	No	No	Yes	Yes
Month Fixed Effects	No	No	No	Yes
N	963	963	963	963

Note: The dependent variable is the average compliance (weighted with the RI parameter weights) for each firm-contract-month triplet. Regarding the regressors, $t1$ *is a dummy variable equal to 1 from* $t1$ *onward and zero before then.* $t2, t3+1, t5+1$ *are constructed analogously. We indicate as* $ty + x$ *a breakpoint taking place* x *months after the Acea's announcement date* ty*, where* $y = 1, ..., 5$*. All regressions also control for the Safety share – the weighted average share of safety parameters – and Job type – the proportion of contracts classified as public illumination, – both calculated among those parameters audited in the firm-contract-month triplet. Standard errors in parentheses. * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).*

increase in performance is not driven by changes in the composition of the set of parameters audited or of firms inspected.[22]

Two obvious concerns that may be raised involve how corruption and multitasking might lead to biased audits. We already mentioned, but it is worth recalling, the mechanism Acea used to limit corruption risks. Each week the 12 engineers in the auditors' office were randomly allocated to three-member teams and then the teams were randomly allocated to contracts to be audited in that week. Rotation should help to sever any link between specific suppliers and auditors, while the random composition of the teams reduces the likelihood that a collusive agreement can be formed. Furthermore, the auditors have no direct benefit to their wage or career progression from assigning more positive (or negative)

[22] In the appendix, we show that over time the set of parameters remains identical in terms of the proportion of weights allocated to quality and safety (Figure A.2), despite an increase in the number of audits per month (Figure A.3). Moreover, improvements over time are also evident for each of the 2 indicator classes of quality and safety, Figure A.4, as well as for the individual parameter categories, Figure A.5. Similarly, the result holds across different sets of firms, as seen in Figure 3.

scores to the firms inspected. Clearly, this system also helps to counteract concerns linked to the auditor scoring heterogeneity. The multitasking concern is that performance improves exclusively on the audited tasks, while staying constant or even worsening on dimensions outside the audit process. In the data, however, we observe both cost overruns and delays in contract completion and, although they are not part of the RI, none of them worsens. Acea's engineers explain this fact as the necessary outcome of the 136 parameters being exhaustive of the contract quality/safety features. Further evidence on a limited role for multitasking is presented below when looking at external performance measures from the regulatory data.

Finally, an important question is whether we can consider the improved compliance to be the result of strategic decisions by suppliers to improve their performance. Indeed, in experimental settings, the mere change in the environment might trigger forms of *Hawthorne effect* (or observer effect). Hence, the mere change from paper-based to digitalized audits might have led suppliers to improve their performance.[23] To rule out this possibility, we can compare how the probability of observing a compliant parameter changes between the audits held before and after $t1$. The estimates in Table A.1 in the appendix show that parameters receiving a higher weight in the announced scoring formula pass from being the ones more likely to be non-compliant before $t1$, to being the most likely to be compliant after $t1$. Furthermore, the parameters more likely to be compliant post $t1$ are those that experts consider faster to adjust.[24] These results are indicative of suppliers effectively changing their behavior.[25] These findings are also interesting to rationalize why the evolution of the external performance measures shows a gradual improvement over time after $t1$ and not a sharp jump like that of the internal measure: supplier improvement across parameters was gradual and they improved more promptly on the safety parameters than on the quality ones

[23] The *Hawthorne effect* is a change, typically an improvement, in some aspects of behavior in response to the awareness of being observed, see Levitt and List (2011).

[24] With the help of expert engineers, we created an indicator variable, *quick*, taking the value of 1 if the transition from a score of not compliant to one of compliant can be reasonably achieved within a one month time frame without incurring extraordinary costs. For instance, examples of parameters with *quick* equal to 1 are those involving the adequacy of "personal protection tools" (mostly helmets) or the presence of signs warning of ongoing work nearby. The adequacy of the machinery, instead, is an example of a parameter with quick equal to zero. While clearly arbitrary, this dummy variable is helpful to test the reasonableness of the performance response observed in our data.

[25] In line with this interpretation, is the evidence in Appendix Table A.3. There, exploiting the random timing of the audits, we show that all firms respond to the $t1$ announcement, regardless of whether they were ever audited before $t1$ or not. Thus, ruling out not only a Hawthorne-type effect, but also learning.

(due to the higher weights assigned to the former in the RI formula). But since changes in the number and duration of blackouts likely hinge more on the quality of the suppliers' work than on their adherence to the worksite safety parameters, this contributes to explaining the lack of major discontinuities in the external performance measures at $t1$.

B. What Was the Cost for Acea of the Improved Performance? Answering this question is an essential step in evaluating the reform's effectiveness. To measure the price impact of the improved performance, we will closely follow what we learned above about the timing of the performance increases. In particular, to causally estimate how the initial jump in compliance (associated with the announcement at $t1$) affects winning auction discount, we will employ a difference-in-differences (DD) strategy. The units of analysis are the auctions held by Acea (treated group) and by other utility firms (control group), both recorded in the Auctions Data. We estimate a model analogous to that of equation (6), but with contract-level data:

$$D^w_{ift} = a_f + b_t + cX_{ift} + \beta_{t1}(Treatment) + \epsilon_{ift}, \tag{4}$$

where D^w is the winning discount (over the reserve price) and the index i indicates the auction, f the entity awarding the contract and t the year. $Treatment$ is a dummy variable equal to one for the contracts awarded by Acea from $t1$ onward and zero otherwise. The coefficient of interest is β_{t1}, the effect of the announcement on the winning discount, conditional on fixed effects for the entity awarding the contract (a_f) and time (b_t), and on other covariates (X) involving contract characteristics.[26]

The other break systematically detected by the Bai-Perron test is at $t5 + 1$, when the performance growth slows down. We thus estimate a second DID model with two breaks: one at $t1$ and one at $t5 + 1$ to account for the two differential phases of RI accumulation and stabilization. It also captures the dynamic from Figure 4 in which the sharp rise in discounts

[26] We present estimates with different specifications for X. We always include a dummy for whether the award procedure includes a provision for the automatic exclusion of abnormally low tenders (a common type of provision across Italian public procurement auctions, see Decarolis (2014)). In some specifications, we also add a dummy variable for whether the object of the job involves public illumination works and four dummy variables for levels of the reserve price (below €250 thousand; €250 thousand and €0.5 million; between €0.5 million and €1.5 million; above €1.5 million).

immediately after $t1$ is followed by a reversion to discounts closer to the ones observed for the control group. Thus, we extend the previous model to include a dummy for auctions held from $t5+1$ onward, $D_{t>t5+1}$:

$$D^w_{ift} = a_f + b_t + \beta_{t1} Treatment_{ft} + \beta_{t5+1} Treatment_{ft} * D_{t>t5+1} + D_{t>t5+1} + \gamma X_{ift} + \varepsilon_{ift}, \quad (5)$$

where β_{t1} measures the effect on Acea's award discounts past $t1$, but before $t5+1$, while β_{t5+1} measures the same effect for being after $t5+1$, relative to the $t1$ to $t5+1$ period. Hence, the effect of the RI accumulation phase is captured by β_{t1}, while that of the RI stabilization phase is captured by β_{t5+1}. The identification of the key parameters in the two models above crucially hinges on the validity of the auctions in the control group to capture price variations that would have affected Acea's auctions absent its reform. As discussed earlier, the graphical comparison of the evolution of winning discounts between the treated and control groups supports the fact that the similar pre-$t1$ dynamics in the treatment and control auctions make the parallel trends assumption likely to hold. We proceed by first presenting our baseline DD estimates and then exploring their robustness to both identification and inference concerns.

Table 8: Baseline Price Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
β_1	3.48	3.63	3.53	6.22***	6.26***	5.96***
	(2.82)	(2.76)	(2.70)	(1.88)	(1.80)	(1.81)
β_2				-5.92*	-5.66*	-5.23*
				(2.30)	(2.23)	(2.16)
Reserve Price FE	No	Yes	Yes	No	Yes	Yes
Object & Res.Pr. FE	No	No	Yes	No	No	Yes
N	4577	4577	4577	4577	4577	4577
R^2	0.47	0.47	0.47	0.47	0.47	0.47

Note: the dependent variable is the winning discount. The sample includes auctions by Acea (treatment group) and all other contracting authorities (control group). The first three columns report estimates for the model in equation (4), while the last three columns report estimates for the model in equation (5). For each model, the model specification gradually expands the set of contract characteristics included as controls: award criterion (columns 1 and 4), also fixed effects for four levels of the reserve price (columns 2 and 5) and also a dummy for whether the contract is for public illumination (columns 3 and 6). Standard errors clusters by year and CA are reported in parentheses. * $(p < 0.10)$, ** $(p < 0.05)$, *** $(p < 0.01)$.

Table 8 presents these baseline estimates for the models in equation (4) (first three columns) and in equation (5) (last three columns). For each model, estimates of three specifications differing in the set of covariates, X, are presented: we first include only a control for whether the award rule involves the automatic elimination of abnormally low bids (columns 1 and 4), then we add four dummy variables for the level of the reserve price (columns 2 and 5) and then also a dummy variable for whether the job involves public illumination works. The results in Table 8 show the lack of any price effect when the post-$t1$ period is considered altogether (first three columns). In addition to not being statistically significant, the estimated coefficients are relatively small in magnitude (a 3.5 percent decline), when compared to the major shift in performance documented above. Interestingly, the estimates change, revealing a rich price dynamic if the post-$t1$ period is divided into a phase pre and post $t5 + 1$. The estimates in the last three columns confirm the visual evidence discussed earlier: discounts initially increase, by about 6 percent of the reserve price, but subsequently decline by approximately the same amount. The magnitudes of β_1 and β_2 are indeed statistically the same at the 1 percent confidence level. Across model specifications and samples, all estimates of β_1 are highly statistically significant, while β_2 is less precisely estimated due to the systematically higher standard errors relative to those of β_1. With the control group of auctions held in central Italy, all estimates are qualitatively the same albeit somewhat larger in magnitude.

In the next section, we will discuss a rationale for these results. In a nutshell, the argument will be that, while firms improved their compliance with the performance measures, they also competed more fiercely to win auctions. Only suppliers with ongoing contracts can be scored and accumulate RI points to be used under the forthcoming SR award system. However, as all firms reached a high score, two forces push toward lower discounts: first, they need less to get additional scores and, second, increasing performance when its level is already high is very costly. Hence, discounts declined in this phase. The β_2 estimate indicates a less pronounced decline than that indicated by Figure 4 because the regressions control for auction characteristics.[27] That figure also reveals that winning discounts increased

[27] The control variable driving most of this difference is a dummy for whether the auction is an "average bid auction." This is a form of modified first price auction incentivizing low discounts. See discussion below.

once again during the short period in which SR were introduced, to about 30 percent, and remained relatively high afterwards.[28]

Table 9: Robustness Checks: Contamination and Comparability

Panel (a): No Contracting Authorities in Central Regions

	(1)	(2)	(3)	(4)	(5)	(6)
β_1	3.48	3.69	3.58	6.08^{***}	6.18^{***}	5.88^{***}
	(2.60)	(2.54)	(2.48)	(1.64)	(1.55)	(1.57)
β_2				-5.61^{*}	-5.38^{*}	-4.93^{*}
				(2.30)	(2.24)	(2.17)
N	3726	3726	3726	3726	3726	3726
R^2	0.50	0.50	0.50	0.50	0.50	0.50

Panel (b): Contracting Authorities in Central Regions

	(1)	(2)	(3)	(4)	(5)	(6)
β_1	3.58	3.31	3.33	7.19^{***}	6.67^{***}	6.37^{***}
	(2.68)	(2.68)	(2.56)	(1.57)	(1.59)	(1.58)
β_2				-7.57^{**}	-7.06^{**}	-6.38^{**}
				(2.57)	(2.39)	(2.31)
N	1161	1161	1161	1161	1161	1161
R^2	0.34	0.35	0.36	0.35	0.36	0.37
Reserve Price FE	No	Yes	Yes	No	Yes	Yes
Object & Res.Pr. FE	No	No	Yes	No	No	Yes

Note: the organization of the table is isomorphic to that in Table 8, with the only difference being the control group observations: panel (a) excludes all auctions held by contracting authorities in Central Italy regions; panel (b) includes only auctions held by contracting authorities in Central Italy regions. Thus, the estimates in panel (a) are less likely to be biased by contamination effects, while those in panel (b) are more likely to be based on more comparable contracting authorities. * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

In Table 9, we explore the robustness of the baseline estimates to two concerns: contamination and comparability of the control group observations. The ubiquitous problem in applying methods like DD to industrial organization problems is the trade off between

[28] However, we do not attempt to estimate the causal effect of the introduction and removal of the SR for three reasons. First, since the transition toward higher performance was essentially done by $t5+1$, the following period of nearly one and half year between $t5+1$ and the SR's introduction represents the ideal setting to study the price effects of higher quality without the additional effects produced by implementation of the SR on bidding and entry behavior. Second, we know from interactions with market participants that even before the SR was removed, suppliers held heterogeneous beliefs about its removal and the commitment of Acea to the hybrid price-only mechanism. Third, since discounts increased in this final period, then the cost-effectiveness considerations presented in the next section would be further strengthened by the lower prices experienced by Acea in the long run. The benchmark of a net zero-price effect thus represents a conservative, but reasonable choice.

these two concerns: contracting authorities closer to Acea are more likely to be comparable to it because, for instance, they use the same suppliers or because their suppliers buy inputs in the same markets. But, the same forces enhancing comparability, induce contamination concerns: changes in the suppliers' technology or behavior triggered by the Acea's experiment might alter the prices observed in the auctions of other contracting authorities. To ease these concerns, the estimates in Table 9 partition the control units auctions depending on whether the contracting authority is located outside Central Italy (top panel) or in it. Acea being itself located in Central Italy is likely more comparable to the control units in panel (b), but also less likely to induce contamination in the control units in panel (a). The comparison of Table 9 estimates with the baseline ones reveals qualitatively identical results.

Finally, there are two additional concerns that we explore through different sets of robustness checks. The first concern is the presence in the sample of auctions entailing specific regulatory features in terms of either firms' participation (restricted auctions) or the elimination of abnormally low discounts (average-bid auctions). We thus repeat the baseline estimates by excluding from the sample either restricted or average-bid auctions (see appendix Table A.4). The second concern regards inference. We evaluate potential problems with inference by using alternative methods for standard errors calculation correcting for serial correlation as in Bertrand, Duflo and Mullainathan (2004) or the one-treated unit problem as in Conley and Taber (2011) (see appendix Table A.5).[29] In all cases, the estimates obtained are qualitatively in line with the baseline estimates discussed above.

C. Effects on the external performance measures - We evaluate the impact of the Acea's announcement at $t1$ on the six external performance measures introduced earlier. This is relevant both as an additional check that multitasking effects are not muting the benefits of the reform implied by the internal performance measures and as an assessment of the reform on measures that are highly socially valuable. The estimation strategy is again a

[29] The latter is the fact that the level at which the treatment effectively takes place is that of the procurer and we observe only one procurer, Acea, receiving the treatment. Hence, any shock hitting Acea at $t1$ biases the estimate of β_1. Nevertheless, as explained in the appendix, the presence of a large control group allows us to conduct valid inference under the Conley and Taber (2011) procedure.

DID based on the following equation:

$$O_{ft} = a_f + b_t + cX_{ft} + \beta_0 D^{Acea*Post} + \epsilon_{ft}, \tag{6}$$

where O_{ft} is one of the performance outcomes that we observe at the level of firm, f, and year, t. On the right hand side of the equation, a_f and b_t are fixed effects for firms and years, X_{ft} is a matrix of controls that includes the number of customers and, finally, $D^{Acea*Post}$ is a dummy for Acea's auctions held after 2007. The coefficient of interest is β_0, which thus captures the difference in external performance between Acea and other firms, after Acea announced the change in the adopted award criterion in December 2007.

Table 10: Estimates for the External Performance Measures: Electricity and Water Sectors

	(1) Long-lasting blackouts	(2) Length long-lasting blackouts	(3) Short-lasting blackouts	(4) Programmed power cuts	(5) Length programmed power cuts	(6) WaterLeakage (full sample)	(7) WaterLeakage (above 1m)
β	-0.325**	-43.272***	-0.922***	0.141*	19.839**	-0.003	0.009
	(0.163)	(13.350)	(0.296)	(0.074)	(9.154)	(0.010)	(0.015)
Observations	386	386	298	386	386	253	59
Firm & Year FE	YES	YES	YES	YES	YES	YES	YES
R-squared	0.843	0.574	0.826	0.720	0.788	0.816	0.890
Sample	All	All	All	All	All	All	Reduced

*Note: The table reports the difference-in-difference estimates for the available external performance measures. In the first five columns, the outcomes cover the electricity distribution sector, whereas the last two columns regard the water distribution sector. ACEA is the treated unit and the treatment is the interaction term of indicators for ACEA and post year 2007. The control units for the electricity sector include all the distributors with at least 200 thousand clients. For the water sectors, the control units include either all the distributors (column 6) or only those in charge of geographical areas with at least 1 million customers (column 7). Robust standard errors in parentheses. Significance: *** p<0.01, ** p<0.05, * p<0.10.*

Table 10 reports the estimates. The first five columns cover different measures of the quality of electricity distribution, while the latter two cover water leakages for both the full sample of firms and for the subsample of larger firms. These estimates confirm the graphical evidence provided earlier: for the five outcomes measuring quality in the low-tension electricity distribution sector, the effect of the treatment is to reduce the number and length of long-lasting blackouts, reduce the number of short lasting blackouts and, on the contrary, increase programmed power cuts. The latter is most likely a signal of greater maintenance efforts. For the water sector, where no RI was introduced, Acea did not improve its performance (in terms of leakage) relative to other comparable firms. Regardless of whether we consider all distributors or only the largest players, the finding of no effect remains. Overall, these estimates confirm the presence of a long lasting performance improvements.

V Cost-effectiveness Analysis

We now present a back of the envelope cost-effectiveness analysis comparing outcomes under Acea's reform and under the status quo absent any reform. An exhaustive cost-benefit (or welfare) analysis would require assigning a monetary value to the increased compliance on all quality and safety parameters. In the spirit of the cost-effectiveness approach, we focus on a subset of specific outcomes associated with the experiment, in terms of both quality and safety.

We start from the quality dimension. Here we focus on the quality of the service measured by one of the external measures of performance, the duration of long-lasting blackouts. We thus convert the estimate in column 2 of Table 10 into a measure of the number of hours of blackout avoided per year: 43.272 hours on average per client. In the post reform period, Acea has on average 1,597,066 customers, divided into 1,277,653 residential and 319,413 business customers. From the official statistics of the regulator (Arera),[30] we associate a cost of blackouts of 2.5 euro/hour for residential customers and of 18.75 euro/hour for business customers. The result is that the reduction in blackouts implies a benefit of 6.623 million euro, 39 percent of which accrues to business customers and the rest to residential ones.

Next, we look at the safety dimension. Here we focus on the change in the probability of fatal accidents as implied by improvements in the subset of internal measures covering safety parameters. Construction and maintenance jobs for electricity generation and transmission are among those with the highest incidence of workplace accidents, including deadly accidents.[31] The occurrence of such accidents has costs for both society and Acea, and the

[30] See Arera's decision n. 172/07 of 12/07/2007.

[31] Electricity is widely recognized as a serious workplace hazard, exposing employees to electric shocks, burns, fires, and explosions. A search among local newspapers revealed that 4 workers had died in the last 15 years while performing jobs procured by Acea. In the U.S., the Bureau of Labor Statistics recorded a

public ownership of Acea only increased its management's concern about these safety risks.

To map the relationship between changes in safety parameter compliance and the occurrence of fatal accidents, we use the statistical model used by Acea's engineers which is known as *Heinrich's pyramid* and is often used by practitioners in the context of industrial systems to link accidents of different intensity.[32] The pyramid entails the following ratios: 1 fatal accident to 10 major accidents, to 30 minor accidents, to 600 material damages, and – finally – to 200,000-300,000 small deviations from safe behaviors. If we assume that each case of non-compliance in the safety parameters audited by Acea corresponds to a small deviation in the pyramid, we can estimate a lower bound for the policy benefit of €3-5 million per year. This is calculated as follows: in a typical audit, 33.08 parameters are assessed, 85.3 percent of which are related to safety (see Table 3). There are on average 43 contracts a year, with an average duration of 250 working days (see Table 4). Suppose that the same rate of compliance observed across audits applies to every working day, then the 55 percent improvement in parameter compliance discussed in sub-section A above implies a reduction in about 163,000 small deviations per year. Using the 200,000-300,000 figure from the pyramid, this maps into a reduction in the probability of a fatal accident of 0.54-0.82 per year. Finally, considering an average of 4 workers on the worksite per day and taking the lowest bound of the OECD (2012) estimates of the "value of a statistical life" of €1.62 million per life saved,[33] the estimated benefit ranges between 3.5 and 5.3 million euro/year.[34]

Finally, regarding the cost, the baseline estimates in Table 8 imply no changes in the winning discounts. Since the winning discount in the auctions is the most relevant cost

total of 5,587 fatal electrical injuries between 1992 and 2013, an average of 254 fatal electrical injuries each year. Death was due either to electrocution or to fires caused by electricity, see Campbell and Dini (2015).

[32] See Heinrich (1931), Bird and Germain (1986) and Goodman (2012). See also its usage by modern safety apps: `http://safesiteapp.com/blog/safety/the-safety-triangle-explained/`.

[33] The number of workers present on the worksite was estimated for us by the same expert engineers who estimated the variable *quick* described earlier. The OECD (2012) values are converted to 2007 nominal euro. We shall also remark that our approach is quite conservative because for benefits we have employed the lowest OECD estimate of the value of a statistical life. Using the upper bound of the OECD estimate (€5.3 million), the benefits would be in the range of €11.55 -17.33. Furthermore, our benefit calculation excludes all the additional savings accruing from both reductions in non-fatal accidents associated with better safety practices and all improvements in quality associated with increased compliance on the quality parameters.

[34] This range is not our interval estimate, but the result of using the two bounds of 200,000 and 300,000 small deviations.

feature of the reform,[35] this no-effect on prices obviously implies that the reform was highly cost effective relative to the status quo in terms of both quality and safety. This conclusion is robust to a worst case scenario analysis. For this, we calculate the reform's effect on the average winning discount directly from the descriptive statistics by taking the difference between the pre-announcements average (21.73 percent) and the year 2010 average (11.91 percent).[36] This gives a reduction in the winning discount of 9.82 percentage points. At an average contract value of €516.1 thousand and considering 43 contracts per year, the total yearly cost increase is then €2.18 million. Hence, we can conclude that, even under a worst case scenario, the benefits from adoption would exceed the costs.

VI Discussion: Entry, Selection and Moral Hazard

Learning the drivers of the performance improvement is crucial in order to replicate Acea's successful reform in other settings. The literature offers three motives related to bidder incentives and information for why first price auctions like those used by Acea before the reform might induce poor contract performance: adverse selection, moral hazard and the winner's curse. In our setting, the latter is unlikely to play a major role, as all bidders are experienced contractors repeatedly bidding in auctions for relatively simple contracts. Regarding selection and moral hazard, distinguishing between them is valuable as they can have different implications for how best to design systems to integrate past performance in procurement.[37]

In our setting, performance improvements can derive from either more effort in the execution stage by contractors, or better selection of contractors, or a combination of both. Indeed, Figure 6 presents evidence consistent with both by showing how the cdf of winning

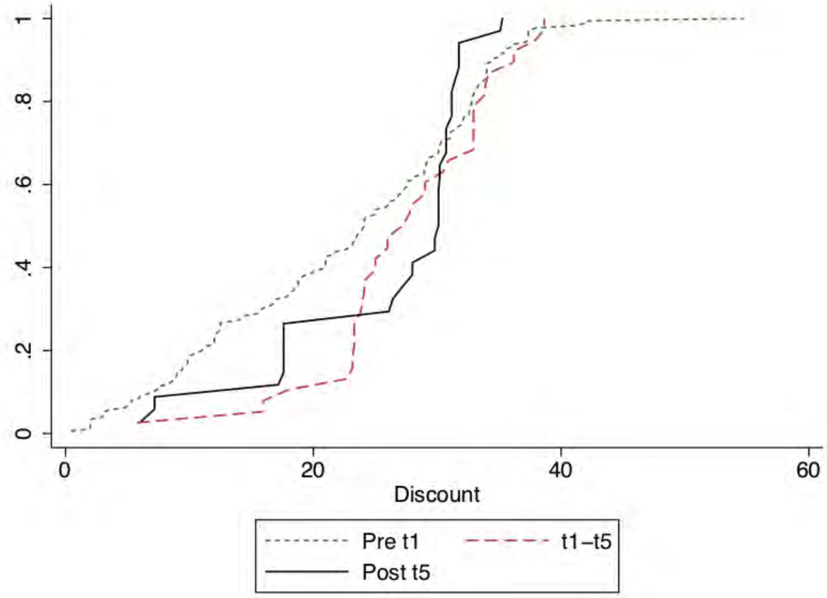
[35] Indeed, according to Acea, carrying out the audits under the new system is no more costly than doing them under the paper-based system.

[36] As shown by Figure 4, using the 2010 average discount as representative of the discount level after the reform is the worst case scenario. If we were to consider the average across the whole period between $t5+1$ and the scoring rule introduction, the level would be higher at 16.19 percent.

[37] For instance, consider the length of the memory of the RI (i.e, how far back should the RI look). This likely needs to be long, possibly infinite, if screening is the concern, but short if moral hazard is predominant; see Elul and Gottardi (2015), and also Kovbasyuk and Spagnolo (2016) who show that optimal memory might differ for positive and negative ratings.

bids in Acea's auctions evolves between those held before $t1$, after $t5+1$, and in between these two periods. The noteworthy aspect is the disappearance post $t1$ of right tail discounts, representing discounts of one third or more relative to the reserve price. It is precisely this type of abnormally high discount that procurers worry will be associated with poor performance. Since replicating Figure 6 for those firms bidding both before and after $t1$ leads to a similar finding of a disappearing right tail after $t1$, we can conclude that the altered bidding behavior of these suppliers is compatible with the presence of moral hazard in contractual performance.

Figure 6: Discount CDF Pre-t1, t1-to-t5 and Post-t5



Note: the plot represents the cdf of the winning bid, dividing bids depending on the timing of the auction: before $t1$, between $t1$ and $t5$ and post $t5$ $(t5+1)$. Source: Auctions data.

Under a moral hazard paradigm, a stylized model of bidding can then rationalizing our earlier findings about the dynamics of winning discounts. In a first price, sealed bid auction, equilibrium bids should depend on two elements: production costs, $C(e)$, which are an increasing function of the effort e that the bidder expects to exert in the execution stage; and a strategic markup, $M(n)$, which is inversely proportional to the number of competitors, n. Prior to $t1$, each auction exists in isolation - the outcome of an auction does not matter for future auctions.[38] Thus, bidders will choose low effort levels to reduce their cost and maximize their profits. However, from $t1$, even if the award rule remains the lowest price, the game played by the contractors becomes dynamic: winning an auction has the additional advantage of potentially being audited and, hence, an opportunity to modify one's own RI while reducing one's rivals' chances of improving their RI. This likely implies changes to both components of the bid relative to the pre-$t1$ case: if better compliance requires more effort, then the optimal $C(e)$ will likely be higher. Moreover, the strategic markup now depends not only on n, but also on the distribution of RI across bidders.[39] Finally, the bid now also incorporates a third element: the continuation value associated with the evolution of the RI. Indeed, winning today and earning a good RI is expected to produce savings in the stream of future auctions, once the scoring rule auction is introduced. This continuation value increases the value of winning today and, hence, balances increases in production costs. Clearly, the relevance of the continuation value depends on how many auctions suppliers perceive they will be able to use their good RI for.

It is not a priori obvious how increasing the RI weight in the scoring auction would affect the outcomes. An increase in this weight helps with the moral hazard problem as it bolsters the benefits of more effort. However, the effect on bidding during the phase before the introduction of the scoring rule is ambiguous. There are two effects, which in a sense correspond to a marginal and an inframarginal effect (alternatively, the intensive and extensive margin). First, winning lowest price auctions gives bidders the opportunity to prove themselves and thus increase their RI (marginal effect). Second, if the implementation of the scoring rule auction is sufficiently delayed that all bidders have the potential to earn a good RI, then symmetric competition in the (future) scoring rule auction will imply that many of the rents from a good RI will be competed out, to the point that winning in the (present) lowest price auctions becomes less attractive (the inframarginal effect).

[38] This assumes that there are no links through, for instance, capacity constraints. This is likely to be a good approximation, since the institutional environment allows for an extensive use of subcontracts that can relax capacity constraint. See Branzoli and Decarolis (2015) for subcontracting and capacity constraints.

[39] That is, even if before $t1$ the environment could be characterized as a symmetric auction with bidders being ex ante identical in terms of costs, after $t1$ firms became asymmetric in terms of their RI stock. This asymmetry can potentially cause changes in the size of the equilibrium markups, for instance by making bidders with lower (or no) RI more willing to shade less their true cost.

Thus, an explanation for the patterns observed in the data is that, right after $t1$, the increase in $C(e)$ was dominated by the changes in the strategic markup and the continuation value. After contractors accumulated a good RI, however, the value of winning an auction in the pre-scoring rule period declines as obtaining positive audit reports cannot offer a competitive edge over rivals. Thus, in this phase the increased production cost dominates.[40]

In the data, however, the effects of the new system concerned not only bidding and performance, but also participation choices. Indeed, while the summary statistics show that the number of bids submitted remains stable and approximately equal to 11 both before and after $t1$, the set of bidders changed in Acea's auctions: while there are 34 suppliers placing at least one bid both before and after $t1$, there are other 36 suppliers who place at least one bid before $t1$, but no bid afterwards. We refer to the latter group of firms as "*exiters*" and to the former as "*stayers*." There are also 3 new entrants placing bids only after $t1$, but never before then. This implies that the average number of bids placed per bidder doubles; from 0.16 (i.e, 11/70) to 0.30 (i.e., 11/37). This increased participation is due to the *stayers*, not to unusually high bidding frequencies for the 3 new entrants and is likely driven by the same incentive to earn RI that we discussed when analyzing the evidence on winning bids. As regards *exiters* and new entrants, however, their mere presence potentially indicates that the experiment might have also triggered some selection effects.

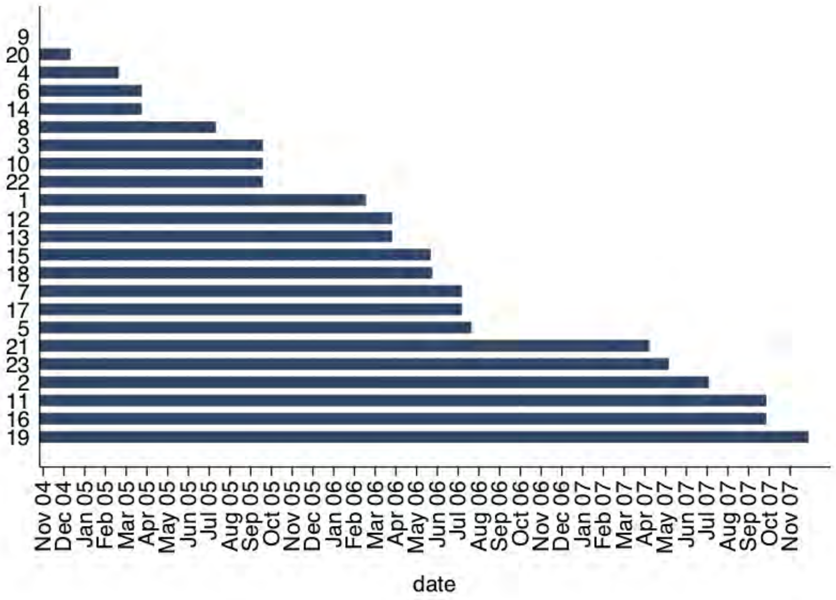
If we focus on *exiters*, however, the data provides only weak evidence of possible selection effects.[41] In particular, Figure 7 shows the timing of the exits does not seem clearly linked to $t1$. This figure reports the last date at which each of the *exiters* (represented by the numerical identifiers on the vertical axis) placed a bid. The smooth path of exits indicates

[40] The intuition for this latter effect is that higher effort pre-scoring rule improves a bidder's expected payoff once the scoring rule becomes effective. However, in a symmetric equilibrium, all bidders win the same number of lowest price auctions and assign the same value to effort. This implies that the equilibrium payoff once the scoring rule is implemented is independent of the weight it assigns to the RI relative to price; the only effect of increasing this weight is thus to increase effort early on. But an increase in this effort decreases the expected payoff from winning an auction pre scoring rule. Finally, this leads bidders uniformly to bid less aggressively in the pre scoring rule period. This result bears some resemblance to models where the strategic effect of an exogenous change (in this case, the expected change in the weight assigned to the RI from zero to 25 percent) more than outweighs any positive direct effect, to the point that equilibrium payoffs are decreasing in the RI weight (see Cabral and Villas-Boas (2005)). We are grateful to Luis Cabral for helping us to elucidate this unintuitive and important element of the strategic environment.

[41] For the 3 new entrants, the type of analysis performed below for the *exiters* cannot be replicated as only one of them could be matched to the firm registry described below.

more of a gradual process than a sharp drop at $t1$.

Figure 7: Last auction date participated (*exiters*)

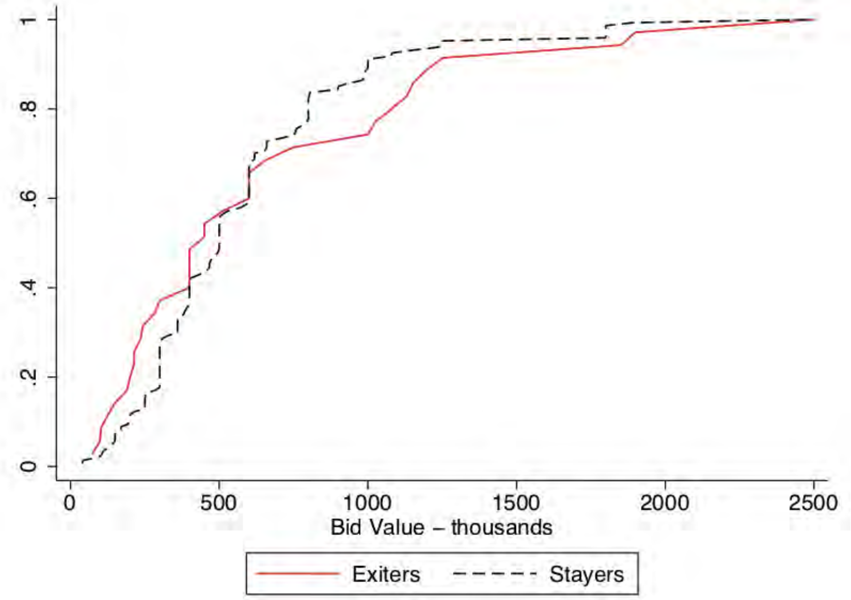


Note: each bar represents the time until the supplier last bids. The figure is drawn for the sample of exiters only. The numbers appearing on the vertical axis are anonymized identifiers of the different firms. Source: Auctions data.

Furthermore, as illustrated by Figure 8, if we compare the cdf of winning bids by both *exiters* and *stayers* (in the pre-$t1$ auctions), we do not observe significant differences. Finally, even in terms of characteristics, *exiters* do not seem to be substantially different from *stayers*. In the appendix, Table A.6 reports summary statistics for the subset of *exiters* and *stayers* that we could match to the Infocamere database, the Italian firm registry.[42] Along most dimensions, *exiters* are smaller than *stayers*; this is the case for revenues, profits and capital. The average number of employees is also lower, but in this case the median is nearly identical. For both groups, the wide variation in characteristics among firms means that the differences in the averages are not statistically significant and it is not obvious how to interpret the results. Thus, to benchmark them we present in panel (b) the analogous statistics obtained for the suppliers active in the auctions of the multi-utility company of the city of Turin. This is the multi-utility company that awards most contracts within the DD control group. Analogously to what was done for Acea, we partition its suppliers into

[42] The registry covers nearly all Italian firms; for a description see Conley and Decarolis (2016).

Figure 8: Bid CDF for exiting and incumbent firms



Note: the plot represents the cdf of winning bids for both exiters and stayers (in the pre-$t1$ auctions). Source: Auctions data.

those bidding both before and after $t1$ (*stayers*) and those bidding only before $t1$ (*exiters*). The comparison of the two groups leads to similar conclusions to those found for Acea's suppliers: the average revenues, profits and capitals are higher among *stayers*. But the data are again characterized by many extreme observations and the result is reversed for revenues and profits when looking at the median.

We conclude that, overall, there is no strong evidence that the pool of *exiters* in Acea's auctions is selected in any particular way relative to the typical exit behavior in the market. Thus, the effects that we uncovered in the earlier sections are likely driven to a large extent by changes in the behavior of Acea's contractors. Perhaps, this is not surprising given that the slow switch to scoring rule auctions allowed most contractors to undertake the steps needed to improve their contractual performance. In turn, this might also be part of the reason why competition remained high in Acea's auctions, thus contributing to limit price increases. In the light of this conclusion, it is interesting to note that most of the policy debate referenced earlier in this paper focuses around issues of supplier selection and most often ignores the potential disciplining effects of reputational mechanisms on moral hazard.

VII Conclusions

This paper has studied the merits of using past performance to spur greater efforts from contractors when executing public works. The evaluation of the evidence from an experiment undertaken by Acea, a large utility company, has shown strong improvements in both the safety and quality of the works after Acea announced its intention to use past performance scores to award future contracts. To some extent this may resemble the well-known *Hawthorne effect*. However, contrary to the *Hawthorne effect*, the improvement was not short-lived, despite the fact that the contractors could have stopped trusting Acea over the delayed implementation of the new award rule, and that it was easier for contractors to improve their score when the starting point was lower rather than later, when the marginal cost of improving increased. Improvements involve all parameters and suppliers, are long-lasting (for at least 10 years after the initial experiment) and are reflected in higher service quality by the utility. Regarding prices, we find some evidence of an initial drop in prices followed by a moderate price increase. Overall, price effects appear negligible when compared to the substantial improvement in performance, as confirmed by a cost effectiveness analysis involving both the duration of blackouts (quality) and the incidence of deadly accidents (for safety). We argue that these results can all be explained by the fact that a reputational mechanism based on objective past performance can effectively curtail supplier moral hazard.

The empirical evidence in this paper points at the very large benefits from imple-menting reputation mechanisms in public procurement for the government and tax-payers. An extension we plan to pursue in future work would involve exploring in greater depth overall welfare implications, which will require building and estimating a structural model. Furthermore, although several different mechanisms might explain how suppliers changed their behavior by increasing their quality and safety performance, it is interesting to note that the explanation offered by the management of Acea is that most of the gains came from improvements in management practices within contractors. Thanks to new data on management practices collected in the last ten years through the World Management Survey (Bloom et al., 2014),[43] there has been increased attention on the role of management in explaining productivity dif-

[43] See: `http://worldmanagementsurvey.org`.

ferences. In this respect, exploring the details of the managerial changes implemented by the suppliers would be useful to understand how (the announcement of) new procurement rules triggered an improvement in management. Regarding this, it is also important to highlight that, while we have stressed the public procurement implications of our analysis, our findings are also relevant to private procurement practices, where the use of vendor rating systems is widespread, but little is known about their effectiveness.

Once the merits of this kind of reputation mechanism in improving contractor performance are proven, many aspects remain open and offer room for future research; for example, how to optimize the parameter weights, how to discipline the rating for new entrants, how to structure the weights in the award criteria, and how to choose the optimal "memory" of the indicator (i.e. how long should be the window of time over which the RI is calculated, and how heavily should older information be discounted). Even the ideal speed at which the switch to a reputation system should occur is an interesting, but little studied problem.

Finally, we conclude by stressing the policy relevance of our findings. There is an ongoing policy debate in both Europe and the US on the use of the past performance of contractors in public procurement. In the US, with the Federal Acquisitions Streamlining Act of 1994, federal agencies started to record past contractor performance evaluations and to share them through common platforms for use in future contractor selection.[44] Interestingly, the EU follows a very different system, essentially barring the use of past performance with the exception of extremely severe types of misbehaviour sanctioned by the judiciary (Gordon and Racca, 2014). Indeed, the use of reputational mechanisms based on past performance has been one of the most contentious issues in the debate leading up to the 2004 and 2014 EU Procurement Directives.[45] To this debate, our results offer a clear empirical illustration of the potential benefits of a reputational mechanism based on objective and clearly targeted past performance measures.

[44] The reform was pushed by Steven Kelman when he served as Administrator of the Office of Federal Procurement Policy in the Office of Management and Budget from 1993 to 1997, playing a lead role in the Administration's "reinventing government" effort that led, among other things, to the Federal Acquisition Streamlining Act of 1994 and the Federal Acquisition Reform Act 1995, see Kelman (1990).

[45] Curiously enough, current EU regulation acknowledges the importance of reputation for some types of procurement. For example, the European Research Council (ERC) funds research (including this study) through peer review, and the track record of the principal invetigator is one of the main selection criteria.

References

Bajari, Patrick, and Steven Tadelis. 2001. "Incentives versus Transaction Costs: A Theory of Procurement Contracts." *RAND Journal of Economics*, 32(3): 387–407.

Bajari, Patrick, Robert S. McMillan, and Steven Tadelis. 2009. "Auctions Versus Negotiations in Procurement: An Empirical Analysis." *Journal of Law, Economics and Organization*, 25(2): 372–399.

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

Banerjee, Abhijit V., and Esther Duflo. 2000. "Reputation Effects and the Limits of Contracting: A Study of the Indian Software Industry*." *The Quarterly Journal of Economics*, 115(3): 989–1017.

Banfield, Edward C. 1975. "Corruption as a Feature of Governmental Organization." *Journal of Law and Economics*, 18(3): 3.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics*, 119(1): 249–275.

Bird, Frank E, and George L Germain. 1986. *Practical loss control leadership.* Loganville, Ga. Institute Publications.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, Daniela Scur, and John Van Reenen. 2014. "The New Empirical Economics Of Management." *Journal of the European Economic Association*, 12(4): 835–876.

Brandon, Alec, Paul J. Ferraro, John A. List, Robert D. Metcalfe, Michael K. Price, and Florian Rundhammer. 2017. "Do the Effects of Social Nudges Persist? Theory and Evidence from 38 Natural Experiments." *NBER Working Paper*, wp23277: 271–280.

Branzoli, Nicola, and Francesco Decarolis. 2015. "Entry and Subcontracting in Public Procurement Auctions." *Management Science*, 61(12): 2945 – 2962.

Burguet, Roberto, Juan Jose Ganuza, and Ester Hauk. 2012. "Limited Liability and Mechanism Design in Procurement." *Games and Economic Behavio*, 76(1): 15–25.

Cabral, Luis M. B., and Miguel Villas-Boas. 2005. "Bertrand Supertraps." *Management Science*, 51(4): 599–613.

Calzolari, Giacomo, and Giancarlo Spagnolo. 2009. "Relational Contracts and Competitive Screening." CEPR Discussion Papers 7434.

Campbell, Richard B., and David A. Dini. 2015. "Occupational Injuries from Electrical Shock and Arc Flash Events." Fire Protection Research Foundation Final Report 1.

Carril, Ricardo. 2019. "Rules Versus Discretion in Public Procurement." *Working Paper*.

Chassang, Sylvain, and Juan Ortner. 2016. "Collusion in Auctions with Constrained Bids: Theory and Evidence from Public Procurement." *Working Paper*.

Colonnelli, Emanuele, and Mounu Prem. 2017. "Corruption and Firms: Evidence from Randomized Audits in Brazil."

Conley, Timothy G., and Christopher R. Taber. 2011. "Inference with Difference in Differences with a Small Number of Policy Changes." *The Review of Economics and Statistics*, 93(1): 113–125.

Conley, Timothy G., and Francesco Decarolis. 2016. "Detecting Bidders Groups in Collusive Auctions." *American Economic Journal: Microeconomics*, 8(2): 1–38.

Coviello, Decio, Andrea Guglielmo, and Giancarlo Spagnolo. 2016. "The Effect of Discretion on Procurement Performance." *Management Science*, Forthcoming.

Coviello, Decio, Andrea Guglielmo, and Giancarlo Spagnolo. 2017. "The effect of discretion on procurement performance." *Management Science*, 64(2): 715–738.

Decarolis, Francesco. 2014. “Awarding Price, Contract Performance and Bids Screening: Evidence from Procurement Auctions.” *American Economic Journal: Applied Economics*, 6(1): 108–132.

Decarolis, Francesco, Raymond Fisman, Paolo Pinotti, and Silvia Vannutelli. 2020. “Rules, Discretion, and Corruption in Procurement: Evidence from Italian Government Contracting.” *mimeo.*

Djankov, Simeon, Oliver Hart, Caralee McLiesh, and Andrei Shleifer. 2008. “Debt Enforcement around the World.” *Journal of Political Economy*, 116(6): 1105–1149.

Doni, Nicola. 2006. “The Importance Of Reputation In Awarding Public Contracts.” *Annals of Public and Cooperative Economics*, 77(4): 401–429.

Elul, Ronel, and Piero Gottardi. 2015. “Bankruptcy: Is It Enough to Forgive or Must We Also Forget?” *American Economic Journal: Microeconomics*, 7(4): 294–338.

Giacomelli, Silvia, and Carlo Menon. 2016. “Does weak contract enforcement affect firm size? Evidence from the neighbours court.” *Journal of Economic Geography*, 17(6): 1251–1282.

Goodman, William M. 2012. “Measuring Changes in the Distribution of Incident-Outcome Severities: A Tool for Safety Management.” *Case Studies in Business, Industry and Government Statistics*, 5(1): 32–43.

Gordon, Daniel I., and Gabriella M. Racca. 2014. “Integrity Challenges in the EU and U.S. Procurement Systems.” *Integrity and Efficiency in Sustainable Public Contracts. Corruption, Conflict of Interest, Favoritism and Inclusion of Non-Economic Criteria in Public Contracts, Gabriella M. Racca & Christopher R. Yukins, eds., Bruylant.*

Heinrich, Herbert William. 1931. *Industrial accident prevention: a scientific approach.* McGraw-Hill.

Hiriart, Yolande, David Martimort, and Jerome Pouyet. 2004. “On the Optimal Use of Ex Ante Regulation and Ex Post Liability.” *Economics Letters*, 84(2): 231–235.

Jofre-Bonet, Mireia, and Martin Pesendorfer. 2003. "Estimation of a Dynamic Auction Game." *Econometrica*, 71(5): 1443–1489.

Kang, Karam, and Robert A. Miller. 2019. "Winning by Default: Why is There So Little Competition in Government Procurement?" Mimeo.

Kelman, Steven. 1990. *Procurement and Public Management: The Fear of Discretion and the Quality of Government Performance.* AEI Studies.

Kim, In-Gyu. 1998. "A model of selective tendering: Does bidding competition deter opportunism by contractors?" *The Quarterly Review of Economics and Finance*, 38(4).

Kolstad, Charles D., Thomas S. Ulen, and Gary V. Johnson. 1990. "Ex Post Liability for Harm vs. Ex Ante Safety Regulation: Substitutes or Complements?" *The American Economic Review*, 80(4): 888–901.

Kovbasyuk, Sergei, and Giancarlo Spagnolo. 2016. "Memory and Markets." *mimeo.*

Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review*, 90(5): 1346–1361.

Levitt, Steven D., and John A. List. 2011. "Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments." *American Economic Journal: Applied Economics*, 3: 224–238.

Lewis-Faupel, Sean, Yusuf Neggers, Benjamin A. Olken, and Rohini Pande. 2016. "Can Electronic Procurement Improve Infrastructure Provision? Evidence from Public Works in India and Indonesia." *American Economic Journal: Economic Policy*, Forthcoming.

Liebman, Jeffrey B., and Neale Mahoney. 2016. "Do Expiring Budgets Lead to Wasteful Year-End Spending? Evidence from Federal Procurement." *Mimeo.*

List, John A., and David Reiley. 2008. "The New Palgrave Dictionary of Economics,." In *Field Experiments.* , ed. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan.

Manelli, Alejandro M, and Daniel R Vincent. 1995. "Optimal Procurement Mechanisms." *Econometrica*, 63(3): 591–620.

Marion, Justin. 2017. "Affirmative Action Exemptions and Capacity Constrained Firms." *American Economic Journal: Economic Policy*, 9(3): 377–407.

OECD. 2012. *Mortality Risk Valuation in Environment, Health and Transport Policies.* OECD Publishing.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy*, 115(2).

Rose-Ackerman, Susan. 1991. "Regulation and the Law of Torts." *The American Economic Review*, 81(2): 54–58.

Shavell, Steven. 1984. "A Model of the Optimal Use of Liability and Safety Regulation." *The RAND Journal of Economics*, 15(2): 271–280.

Spulber, Daniel F. 1990. "Auctions and Contract Enforcement." *Journal of Law, Economics and Organization*, 6(2): 325–44.

Tadelis, Steven. 2016. " The Economics of Reputation and Feedback Systems in E-Commerce Marketplaces." *IEEE Internet Computing*, 20(1): 12–19.

Zheng, Charles Z. 2001. "High Bids and Broke Winners." *Journal Economic Theory*, 100: 129–171.

For Publication on the Authors' Web Page

Web Appendix

I Data

The data used in the paper come from three main sources, plus several ancillary ones. The Audit data come directly from the firm implementing the experiment, Acea (https://www.gruppo.acea.it/en). They were released to us for research and study purposes. The Auction data come from the database on public works of a private company, http://www.telemat.it/. This is a major information entrepreneur (IE) and its main activity is selling information about public contracts to construction firms. For the subset of auctions held by Acea, we also have the internal Acea's records regarding these auctions. The Regulatory Reports data come from the public authority the yearly reports of the Italian Regulatory Authority for Energy, Networks and Environment (ARERA, https://www.autorita.energia.it/it/inglese/). Additional data were obtained from the Observatory on Public Contracts of the Italian Anticorruption Authority http://www.anac.it, from which we take the data on time delays and cost overruns in contract execution. Furthermore, for the cost effectiveness analysis, the value of a statistical life figures come from the OECD (https://www.oecd.org/environment/mortalityriskvaluationinenvironmenthealthandtransportpolicies.htm), while those for the economic cost of 1 hour of blackout, separately for business and residential customers come from Table 11 in the AREA's decision n. 172/07 of 12/07/2007.

II Additional Results

In this appendix section, we present a series of additional results supplementing the various analyses presented in the main text.

- The estimates in Table A.1 explore the behavior of suppliers when they become aware of the new scoring auction. We do so by focusing on the audit data in the period before the introduction of the scoring rule and further partitioning this sample into two subsamples: audits held before and after $t1$. For each of these subsamples, we estimate a series of probit regressions performed at the level of each individual audited parameter. We estimate the following probit model for the probability of the score being 1 (i.e., compliant) on features of parameters, contracts and suppliers:

$$Pr(compliant) = \Phi[t + f + \alpha\ weight + \theta\ quick + \gamma_j \sum_{j=2}^{12} category_j], \qquad (7)$$

where Φ is the normal cdf, *compliant* is the score (0 or 1) taken by the parameter audited, t and f are fixed effects for the year and contractor, *weight* is the weight associated with the parameter, *quick* is a dummy for whether the parameter can be adjusted within one month at a small cost and $category_j$ are dummies for the category to which the parameter belongs.

We are particularly interested in the coefficient on *weight* as this has the potential to reveal the strategic nature of supplier responses. Table A.1 shows the probit marginal effects for two separate samples: audits held in the period before $t1$ (first four columns), and audits held after then (last four columns). We find that the sign of the coefficient on *weight* changes from negative to positive. Thus, after $t1$, suppliers become more compliant in those parameters with the strongest potential to bolster their RI. This switch in the coefficient sign is evident across all specifications, as we move from a baseline model, controlling only for *weight*, and we expand the model to incorporate parameter, contract and firm features.[46]

Regarding the other coefficients in Table A.1, the one on *quick* is useful to assess the potential for collusion between suppliers and monitors. Indeed, performance might be improving because the repeated interaction allows the parties to learn how to collude under the new system. However, this interpretation of the data would seem less plausible if the improvements were concentrated on those parameters that should be faster to effectively adjust. With the help of expert engineers, we created a dummy

[46] All estimates in Table A.1 are based on the subset of parameters that are audited at least once both before and after $t1$. The results remain qualitatively the same for the post-$t1$ sample if all audits are included.

variable, *quick*, that is equal to 1 if the transition from a score of not compliant to one of compliant can be reasonably achieved within a one month time frame without incurring extraordinary costs. For instance, examples of parameters with *quick* equal to 1 are those involving the adequacy of "personal protection tools" (mostly helmets) or the presence of signs warning of ongoing works nearby. Instead, the adequacy of the machinery is an example of a parameter with *quick* equal to zero. While clearly arbitrary, this dummy variable is helpful to test the reasonableness of the performance response observed in our data. Indeed, the finding that the coefficient on *quick* is positive (and that its significance increases post $t1$) is suggestive of suppliers effectively changing their behavior. This interpretation is further strengthened by what we report below with regard to the behavior in the auctions. However, it is relevant here that while it is impossible to fully rule out the possibility of collusion/corruption, the system of random rotation of auditors and of random selection of the sites to inspect was explicitly meant to curtail these types of risks. Indeed, Acea never expressed to us concerns about episodes of corruption or collusion during the period our data cover.

- In Table A.2, we complement the Bai-Perron tests in Table 6 with a series of Chow tests for the presence of one break at $t1$ (odd numbered columns) and five breaks, at $t1, ..., t5$ (even numbered columns). For all six tests, we reject the null of no breaks in favor of the alternative of breaks at the specified dates.

- In Table A.3, we explore a different way to look at the heterogeneity across firms in the announcement response is to exploit the audit randomness. Since at every point in time the choice of which contract to audit is random, we can compare whether the compliance in the very first audit that a firm receives post $t1$ is different between firms audited and those not audited prior to $t1$. In the data, 33 firms receive at least 1 audit post $t1$, with 26 of these audited only post $t1$ and 7 also having already been audited before then. Table A.3 reports the results of linear regressions of the share of compliant parameters in the first audit (post $t1$) on an indicator of whether the firm was already audited prior to $t1$ and other controls.[47] Both when compliance is unweighted (first

[47] The regressions include controls for the share of safety parameters among those audited (in models

three columns) and when it is weighted through the parameters' weight in the RI formula (last three columns), the results indicate a lack of any statistically significant difference between the two supplier groups. Although the small sample size requires interpreting this evidence with caution, this result is also in line with the saliency of the first announcement highlighted by the earlier findings.

- Table A.4 presents robustness checks for the baseline DD estimates involving the award process specifications. All sample auctions share many characteristics, but there are nevertheless subtleties in the regulations defining the precise mechanisms for the contract award that might affect outcomes. Across auctions, differences in both auction procedures and awarding methods exist. Auctions where a restricted set of bidders is invited to bid can be used under certain conditions, and indeed this method is used for 87 out of the 330 auctions held by Acea.[48] Panel (a) reports estimates excluding these 87 auctions. Regarding the award criterion, 42 out of the 330 auctions are awarded via modifications of the lowest price rule. All modifications entail automatically eliminating abnormally low bids (i.e., discounts considered "too good to be true").[49] Panel (b) eliminates from the sample all Acea's auction run with the automatic elimination of the lowest bids. For both panel (a) and (b), the estimates are qualitatively similar to those in the baseline regressions.

- In Table A.5, we evaluate potential problems with inference by using alternative methods for standard errors. The four columns report 95 percent confidence interval estimates corresponding to models (2), (3), (5) and (6) of Table 8. The rows indicating

(2),(3),(5) and (6)), a dummy for whether the contract is for public illumination (in models (2),(3),(5) and (6)), fixed effects for the quarter of the year (in models (2),(3),(5) and (6)) or dummy variables for being past each of the breaks $t2, ..., t5$ (in models (3) and (6)) and to account for the increasing compliance over time (in models (2),(3),(5) and (6)).

[48] The Code refers to these auctions based on invitations as "negotiated procedures." They are studied in Coviello, Guglielmo and Spagnolo (2016).

[49] Acea used the flexibility given to it by the Code to experiment with three alternatives to the lowest price rule. One method entailed awarding the contract to the contractor with the discount closest to the average discount offered, increased by 20 percent. A second method entailed using a trim mean disregrading 10 percent of the highest and lowest discounts (Decarolis, 2014). The third method entailed randomly deciding after the bids were submitted whether the criterion to be used was the highest discount or the discount closest to an average of the submitted discounts (either their simple average or their trim mean). The award criterion is always specified in the call for tenders, so bidders knew these 42 auctions were different and this might have altered their bidding.

"CA-Year" report estimates where the clustering is at the year and CA level, as in Table 8. The other rows in the table present two alternatives. The rows "CA" use clustering at the CA level only. As is well known from Bertrand, Duflo and Mullainathan (2004), this can serve to correct for overestimation of the significance of the treatment effect driven by autocorrelation in the data. The table reveals that this correction has no qualitative implications for our results: relative to the baseline estimates, for all models involving both β_1 and β_2 there are no changes, while significance increases for models involving β_1 only. The latter models indicate a negative and significant effect at the 95 percent level. Since the clustering at CA and year is preferable to account for time variation, however, we prefer to rely on our more conservative baseline estimates. The second concern regarding inference is the fact that the level at which the treatment effectively takes place is that of the procurer and we observe only one procurer, Acea, receiving the treatment. Hence, any shock hitting Acea at $t1$ biases the estimate of β_1. As argued by Conley and Taber (2011), if the shocks potentially hitting Acea and the control CAs belong to the same distribution, and if a sufficiently large number of control CAs are observed, valid inference can still be conducted by adjusting the standard errors. Since we have many control CAs, we use the Conley and Taber (2011) method to assess how significance changes relative to our baseline estimates. The "Conley-Taber" rows are indeed different from the baseline ones: in columns (3) and (4), β_1 loses significance, while β_2 loses significance in model (4) when the largest set of control CAs is used. Overall, this indicates that we should be cautious in interpreting the findings in Table 8 about significant and opposite signs of β_1 and β_2. Hence, as before, a more conservative interpretation is that there are no statistically significant price changes throughout the sample.

- In Table A.6, reports balance sheet summary statistics from Infocamere, the registry of Italian firms. The data is reported separately for *exiters* and *stayers*.

- In Figure A.1, we present the evolution over time of additional external performance measures: the number and duration of planned power cuts.

- Figures A.2, A.3, A.4 and A.5 explore the heterogeneity in the composition of the audits over time. We begin by looking separately at parameters in the quality and safety classes. Figure A.2 reports for each month the total weight (averaged across all audits in the month) of parameters relating to these two classes. Safety parameters always carry a higher total weight, but their proportion relative to the quality parameters remains rather stable over time. Indeed, the evolution of the monthly average weighted parameters in these two classes reported in Figure A.4 confirms a clear upward trend for both of them. As the latter four columns of Table 6 show, breaks in both series occur at $t1$, but the dates of the other breaks are not all identical. This is also related to the speed of adjustments in compliance, as we will discuss below. Before that, we complete the graphical analysis of the composition issue by taking an even more disaggregated view of the performance measures through their grouping into categories. As Figure A.3 shows, the number of parameters audited per month is quite heterogenous, but Figure A.5 reassures us that the compliance increase over time is quite homogenous across categories.[50] Similarly, while there is heterogeneity in how many audits each contractor receives,[51] performance increases are rather homogenous across contractors.

[50] To make the figure easier to interpret, we reported only the 4 most audited categories, but the increase is present essentially in all 12 categories, as also revealed by the summary statistics in Table 3.

[51] Ranging from nearly 200 audits for the most audited contractor to zero audits for a few contractors.

Table A.1: Probability of Compliant Parameter

	Pre-announcement				Post-announcement			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weight	-0.026***	-0.024***	-0.024***	-0.025***	0.011***	0.013***	0.013***	0.013***
	(0.005)	(0.007)	(0.007)	(0.007)	(0.001)	(0.001)	(0.001)	(0.001)
Quick		0.077*	0.077*	0.074*		0.066***	0.065***	0.066***
		(0.036)	(0.036)	(0.036)		(0.006)	(0.006)	(0.006)
C2-Documentation		-0.412***	-0.412***	-0.440***		-0.284***	-0.268***	-0.270***
		(0.053)	(0.053)	(0.055)		(0.010)	(0.010)	(0.010)
C3-Works Execution		-0.518***	-0.518***	-0.523***		-0.189***	-0.189***	-0.192***
		(0.062)	(0.062)	(0.064)		(0.010)	(0.010)	(0.010)
C7-Underground works		-0.302***	-0.302***	-0.296***		-0.291***	-0.288***	-0.286***
		(0.051)	(0.051)	(0.052)		(0.009)	(0.009)	(0.009)
C9-Personnel		-0.308***	-0.308***	-0.332***		-0.349***	-0.359***	-0.365***
		(0.069)	(0.069)	(0.069)		(0.011)	(0.011)	(0.011)
C10-Works site regularity		-0.673***	-0.673***	-0.680***		-0.449***	-0.443***	-0.441***
		(0.054)	(0.054)	(0.056)		(0.009)	(0.009)	(0.009)
C11-Works site safety		-0.381***	-0.381***	-0.405***		-0.272***	-0.272***	-0.275***
		(0.056)	(0.056)	(0.057)		(0.010)	(0.010)	(0.010)
Year Fixed Effects	No	No	Yes	Yes	No	No	Yes	Yes
Firm Fixed Effects	No	No	No	Yes	No	No	No	Yes
N	1,702	1,374	1,374	1,374	56,085	44,653	44,653	44,653

This table reports the marginal effects of probit regressions. The dependent variable is the score on the parameter: 1 if compliant and 0 if not compliant. The first four columns regard the subsample of scores assigned in the audits held before $t1$, while the latter four columns regard audits that occurred after $t1$.

Table A.2: Breakpoints in the Internal Performance Measures (Chow Tests)

Weighted Compliance		*Quality*		*Safety*	
1 break	5 breaks	1 break	5 breaks	1 break	5 breaks
at t1	at t1-5	at t1	at t1-5	at t1	at t1-5

Note: The table reports the results of Chow tests. The variable is the monthly weighted average compliance, measured on all audited parameters (first two columns) or on the subset of quality parameters (next two columns) or safety parameters (latter two columns).

Table A.3: Compliance at the First Audit

	Unweighted			Weighted		
	(1)	(2)	(3)	(4)	(5)	(6)
Audited before t1	-0.205	-0.220	-0.219	-0.208	-0.223	-0.222
	(0.138)	(0.135)	(0.133)	(0.141)	(0.137)	(0.135)
Safety share		0.411	0.464		0.461	0.508
		(0.356)	(0.335)		(0.362)	(0.339)
Object (PI)		-0.160	-0.155		-0.161	-0.155
		(0.124)	(0.124)		(0.126)	(0.125)
N	33	33	33	33	33	33
Quarter FE	Yes	Yes	No	Yes	Yes	No
Breaks	No	No	Yes	No	No	Yes

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The sample consists of the 33 firms audited at least once post $t1$. Out of these 33 firms, 26 firms were never audited pre $t1$ and 7 firms had already been audited before then. The table reports OLS coefficients for regressions of the share of complaint parameters during the first audit on a dummy for whether the firm was already audited pre $t1$, and other controls. The set of controls changes across columns and includes combinations of: quarter fixed effects, the share of safety parameters in the first audit and dummy variables for both the dates of the breaks and whether the contract audited is for public illumination. In the first three columns, the compliance measure is unweighted, while in the latter three it is weighted by the weights in the RI formula.

Table A.4: Robustness Checks: Award Process Specifications

Panel (a): No Restrictions to Open Competition

	(1)	(2)	(3)	(4)	(5)	(6)
β_1	4.75**	4.95**	4.80**	5.40***	5.64***	5.43***
	(1.61)	(1.53)	(1.56)	(1.45)	(1.36)	(1.42)
β_2				-2.21***	-2.32***	-2.14***
				(0.65)	(0.63)	(0.63)
N	4392	4392	4392	4392	4392	4392
R^2	0.48	0.48	0.48	0.48	0.48	0.48

Panel (b): No Variations to the Lowest Price Criterion

	(1)	(2)	(3)	(4)	(5)	(6)
β_1	3.70	3.62	3.43	6.52***	6.46***	6.02***
	(2.51)	(2.53)	(2.38)	(1.36)	(1.38)	(1.39)
β_2				-6.02***	-6.04***	-5.50***
				(0.78)	(0.78)	(0.80)
N	3531	3531	3531	3531	3531	3531
R^2	0.43	0.43	0.43	0.43	0.43	0.44
Reserve Price FE	No	Yes	Yes	No	Yes	Yes
Object & Res.Pr. FE	No	No	Yes	No	No	Yes

This table contains results to evaluate the robustness of the baseline DD estimates in Table 8 with respect to tender specifications. Panel (a) reports estimates excluding auctions with restricted participation. Panel (b) reports estimates excluding auctions where the price-only award criterion involves the automatic elimination of abnormally low tenders.

Table A.5: Robustness: Inference

Panel (a): All Contracting Authorities

	(1)	(2)	(3)	(4)
VARIABLES	W.Discount	W.Discount	W.Discount	W.Discount
PA-Year	(-15.1;5.8)	(-14.9;5.6)	(1.4;8.8)	(1.3;8.4)
PA	(-5.7;-3.6)	(-5.7;-3.7)	(4.2;6.0)	(4.0;5.8)
Conley-Taber	(-7.7;-1.2)	(-7.6;-1.2)	(1.5;7.2)	(1.3;7.0)
PA-Year			(-22.8;-6.2)	(-22.5;-5.8)
PA			(-15.7;-13.4)	(-15.4;-12.9)
Conley-Taber			(-36.7;-9.4)	(-36.0;-9.0)

Panel (b): Contracting Authorities in Central Regions

	(1)	(2)	(3)	(4)
VARIABLES	W.Discount	W.Discount	W.Discount	W.Discount
PA-Year	(-13.1;4.2)	(-13.2;3.7)	(3.1;9.2)	(2.2;8.5)
PA	(-7.1;-1.8)	(-7.5;-2.0)	(4.0;8.2)	(3.3;7.4)
Conley-Taber	(-9.2;-1.5)	(-8.9;-1.8)	(1.0;6.0)	(1.1;5.5)
PA-Year			(-23.0;-9.2)	(-22.3;-8.4)
PA			(-18.8;-13.4)	(-18.1;-12.6)
Conley-Taber			(-21.8;-6.7)	(-21.4;-6.5)

The table reports 95 percent confidence interval estimates for the same regression models presented in columns (2), (3), (5) and (6) of Table 8. The estimates in the three rows use different methods to compute standard errors: the top row uses clustering at the year and CA level and is thus identical to the point estimates in Table 8. The second row uses clustering at the CA level to account for autocorrelation. The third row uses the Conley-Taber adjustment for a small number of treatment units.

Table A.6: Summary stats: Exiting and Incumbent firms

Panel (a): Contractors Entering Acea's Auctions

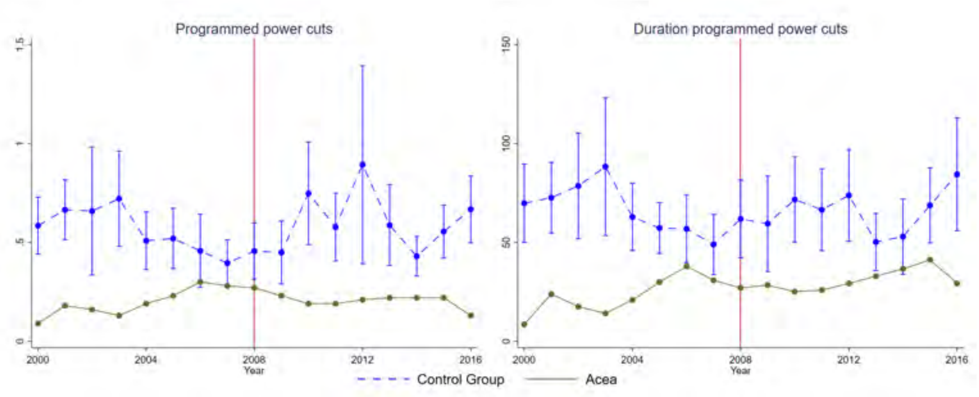
	Exiters				*Stayers*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean	p50	SD	N	Mean	p50	SD	N
Revenues	8,283	2,458	14,615	24	8,934	5,660	9,401	16
Profits	-21	6	697	24	32	5	73	16
Capital	391	36	788	24	998	47	2699	16
Number of Employees	10.3	5	11.1	24	51.7	4.50	180.4	16
Number of Managers	4.96	2	7.57	24	3.38	2	2.55	16
Proportion Female Managers	0.07	0	0.11	24	0.12	0	0.26	16
Public Company	0.96	1	0.21	23	0.88	1	0.34	16

Panel (b): Contractors Entering Turin's IRIDE Auctions

	Exiters				*Stayers*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean	p50	SD	N	Mean	p50	SD	N
Revenues	7,121	4,795	7,127	18	50,860	2,645	152,410	15
Profits	30	15	256	18	736	9.69	2,283	15
Capital	298	40	505	26	10,319	40	43,370	19
Number of Employees	9.04	9.50	5.53	26	15.1	8	15.8	19
Number of Managers	4.35	3	2.96	23	8.11	5	9.45	19
Proportion Female Managers	0.03	0	0.06	26	0.09	0	0.15	19
Public Company	0.71	1	0.46	24	0.72	1	0.46	18

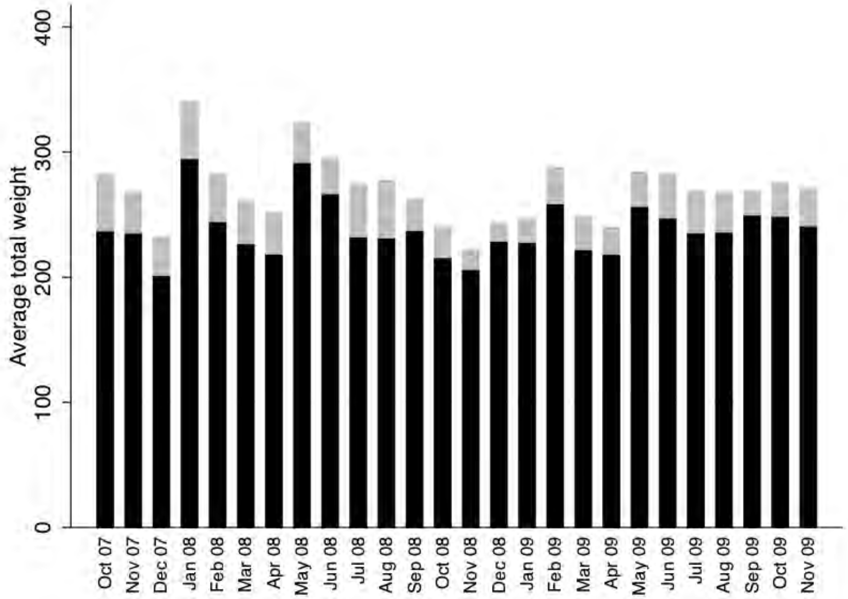
Firm-level summary statistics. Panel (a) refers to the contractors active in Acea's auctions, while panel (b) refers to the contractors bidding in the auctions of Turin's multi-utility company (IRIDE). Across all multi-utilities in the DD control group, this is the one for which we observe most contracts during the sample period. For both Acea and IRIDE, we indicate as *exiters* those contractors observed bidding at least once before $t1$, but never after then, and as *stayers* those bidding at least once both before and after $t1$. For each of the 4 sets, the columns Mean, p50 and SD report the average, median and standard deviation taken across all firms in the set. Column N reports the number of firms considered. Acea characteristics considered are averaged over the years 2006-2010. They are: revenues, profits and capital (all expressed in €1,000), the number of all dependent workers (Number of Employees and Number of Managers), the fraction of female managers over all managers (Proportion of Female Managers) and the share of public companies.

Figure A.1: Evolution of Discounts and External Performance Measures



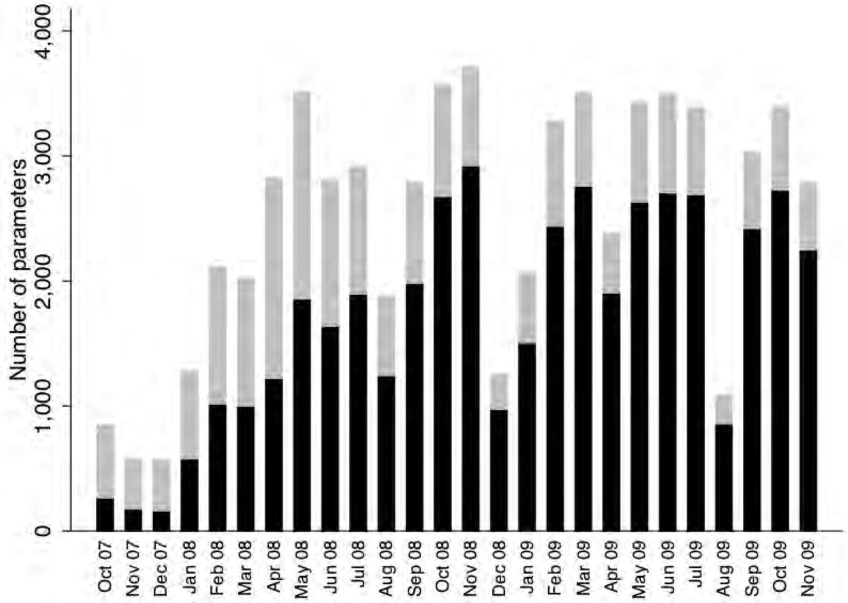
Note: The figure illustrates external performance measures for both Acea (in green) and other providers (in blue). In all graphs, the red, vertical line indicates the $t1$ *announcement date.*

Figure A.2: Safety and Quality: Average Weights across Audits



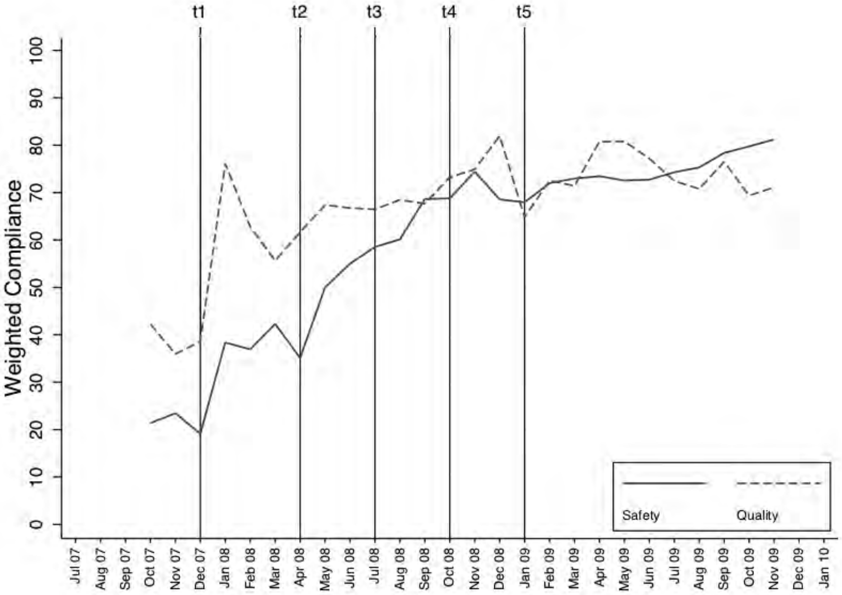
Source: Audits data. The plot represents the total weight, by audit, of parameters relating to Quality dimensions (grey bar) and Safety dimensions (black bar).

Figure A.3: Number of Parameters Audited



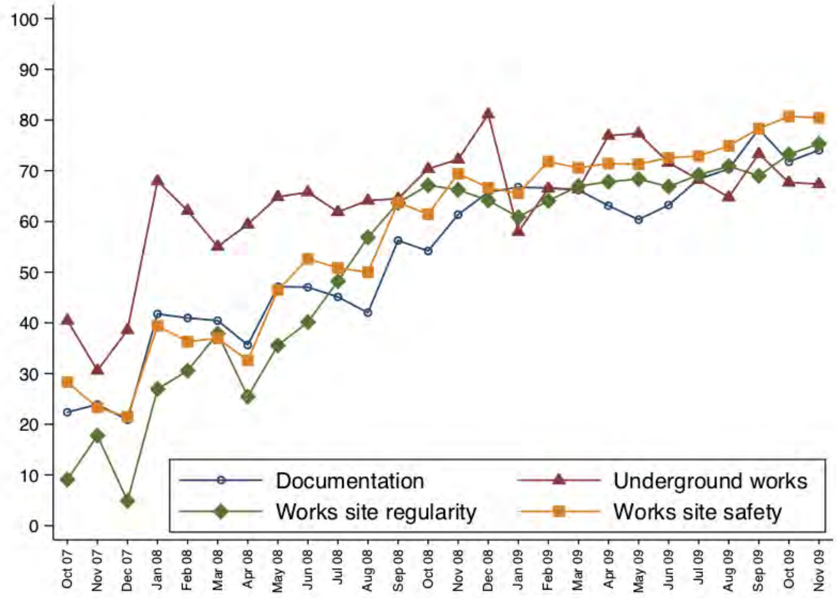
Source: Audits data. The bars represent the total number of parameters checked throughout the month of reference, distinguishing the compliant parameters (in black) from the not compliant ones (in grey).

Figure A.4: Safety and Quality: Evolution of Compliance over Time



Source: Audits data. Monthly average compliance calculated separately for Safety and Quality on all parameters inspected in the month of reference, weighting each parameter by its weight in the RI. The vertical lines identify each announcement date.

Figure A.5: Parameters Audited: Evolution of Compliance over Time



Source: Audits data. The lines show the progress of the reputation index calculated on a monthly basis for each of the four most audited Safety and Quality dimensions.