Past Performance and Procurement Outcomes

Pasha Andreyanov, Francesco Decarolis,
Riccardo Pacini and Giancarlo Spagnolo*

August 12, 2024

Abstract

In procurement, balancing costs with contract execution performance is key. Our study evaluates a unique experiment where scoring auctions were used with price weighed against the quality of past works rather than future ones. Results indicate a significant increase in performance – from 25% to 90% – across all audited parameters, but, paradoxically, no price increase. Using a symmetric auction model, we show that this is possible, especially when the market is characterized by quality concerns. We use structural estimation to quantify this phenomenon.

JEL: H57, D44, D47, K12, C57

Keywords: Public procurement, past performance, audit, vendor rating, public utility

*Andreyanov: HSE University, Moscow, *pandreyanov@gmail.com*; Decarolis: Bocconi University, *francesco.decarolis@unibocconi.it*; Pacini: Agenzia del Demanio, *riccardo.pacini01@agenziademanio.it*; Spagnolo: University of Rome Tor Vergata, EIEF & SITE, *spagnologianca@gmail.com*; We acknowledge financial support from: European Research Council (ERC Starting Grant 679217) to Decarolis; Wallander and Hedelius Foundation (Handelsbanken P2013-0162) to Spagnolo; MUR-PRIN 2017 (2017Y5PJ43) to Andreyanov and Spagnolo.

1 Introduction

In 2023, public spending accounted for approximately 12% of GDP in the United States and 18% in the OECD countries. For the bulk of procurement scenarios, traditional price-only auctions are used. However, intensifying price competition might compromise ex-post performance.[1] As a result, there is a growing demand for more sophisticated evaluation methods that balance price and contract execution performance.

In the public sector, where discretion is limited to minimize corruption risks, this goal can be achieved via a *scoring auction* that weighs price against the quality in the upcoming contract. Such auctions have garnered significant scholarly attention, see Che (1993); Asker and Cantillon (2008, 2010); Lewis and Bajari (2011), among others.[2] However, little is known about scoring auctions where the quality provided in *past* contracts enters the score. Such a system gives an advantage to firms with superior organizational frameworks and reputations, supposedly leading to higher quality and, therefore, higher costs. We will refer to it as the *vendor rating* system.

Our paper studies a pilot run of such a system in a market for construction contracts associated with the electrical grid in a substantial region in central Italy, including Rome. The market is operated by a multi-utility company, Acea, which is largely owned by the municipality of Rome, ensuring that the services provided align with public interests.

Traditionally, Acea awarded these contracts to private firms via *price-only* auctions. However, due to price competition, only the firms with the cheapest possible labor and materials used to win these auctions, leading to repeated blackouts and work accidents. This forced Acea to seek ways to improve its safety and performance standards.

[1] For instance, Spulber (1990) indicates that open competition in the construction sector can instigate adverse selection and ex-post opportunism among contractors. For a contemporary perspective, refer to Lopomo, Persico and Villa (2023) and additional references detailed in the literature review below.

[2] For more theoretical and empirical analysis of scoring auctions, see Andreyanov (2019); Laffont et al. (2020); Camboni et al. (2023). Bid preferences can also be considered a scoring auction version, see Krasnokutskaya and Seim (2011).

In October 2007, Acea's auditors started conducting random worksite audits, meticulously examining the relevant parameters of the job quality and adherence to the needed safety standards in execution. Two months later, Acea informed suppliers that these evaluations would form a *reputation index* (RI). This index would subsequently become a component of a linear scoring rule, attributing 75 percent of the score to pricing and the remaining 25 percent to the RI.[3] The price-only auctions persisted for another 37 months, granting firms enough time to update their RIs before finally switching to scoring auctions in January 2011.

The immediate effect of the switch was that more contracts were being won by firms with higher RI, that is, a better record of past performance. As a result, the safety and quality of current works has considerably improved, and blackouts have become less frequent. Prices, however, did not increase. Furthermore, pressure started to mount from suppliers, seemingly discontented with the approach. First, in a rare episode of unity, suppliers wrote a collective complaint to the mayor of Rome.[4] Second, Acea faced legal contentions over the perceived discriminatory nature of the scoring auction. In a twist of events, Acea shelved the system in May 2011. The auctions resumed in May 2012, but under a *hybrid* system that combined traditional price-only auctions with rigorous compliance checks through audits.

The aftermath suggests that combining the scoring auction with the vendor rating system produced a large increase in performance at no or even a negative additional price for the buyer. Our paper aims to document this empirical puzzle and offer a rational explanation for it.

We used several data sources to assess price and quality changes in the market. First, complete audit data on the parameters measuring quality and safety standards in suppliers' contract execution are observed for 10 years, from the introduction of the new audits in 2007 up until 2017. We refer

[3]It's pertinent to note that Acea's choice of the term "reputation" for its vendor rating system refers strictly to observable, audited parameters. This definition diverges from the traditional economic context where reputation denotes a belief regarding a player's nature Bar-Isaac and Tadelis (2008).

[4]One of the authors, Spagnolo, witnessed this episode.

to these as the *internal measures* of performance. Second, we observe the procurement auctions held between 2004 and 2017 by Acea and every similar utility company in the country. Finally, from the public regulatory authority, we obtained *external measures* about the output quality for both the buyer and comparable utility companies.

We begin the analysis with the events between the first announcement of vendor rating and the switch to the new auction mechanism, to which we refer as the *grace period*. Through this period, compliance in the parameters audited monotonically increased from 25 percent to more than 80 percent just before the switch. 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, see Figure 2.

It is possible, of course, that firms have learned to formally comply with the observed parameters without changing their actual performance. To counter this argument, we study external measures since they are not part of the vendor rating system. By comparing Acea to a control group of utility companies, we find a significant drop in the frequency, duration, and unpredictability of blackouts in the electricity distribution service related to Acea's contracts. At the same time, the quality of the water distribution service provided by Acea – which was not affected by the new procurement system – did not improve or worsen; thus, it is unlikely that the quality improvement was due to a system-wide shock.

Proceeding to the events that took place once the scoring auction was implemented, to which we refer as the *scoring period*, one could expect that the firms, especially those that have raised their RI the most, would take advantage of the new auction mechanism by substantially increasing their bids and that the prices would increase. On the contrary, we observe that the awarding prices have declined, and sharply so. Figure 5 shows that the discounts (i.e., the percentage difference between the price offered and the reserve price set by the buyer) have jumped roughly from 10 to 25 percent. At the same time, the number, the composition of bidders, and their RI remained stable in the neighborhood of the switch.

To explain the puzzle, we employ a static symmetric model of a scoring auction. The model is made intentionally simple to allow for clear identification while tying together all three pieces of evidence: increased quality, decreased prices, and discontent among the bidders. Our model differs from those of Che (1993) and Asker and Cantillon (2010) in two key elements: the costs of building up RI (to which we refer as quality) are sunk, and the types are bivariate. These are both key features of the environment that we analyze, and they reasonably capture a broad range of relevant settings in procurement.

First, as a proof of concept, we show an example where both quality and discounts can increase when switching from the price-only to the scoring auction, see Figure 6. The main ingredient of the example is a positive correlation between cost and quality, which again underscores the presence of quality concerns in the market. This feature is central to the theoretical discourse surrounding adverse selection in contract procurement, as explored by Manelli and Vincent (1995) and Lopomo, Persico and Villa (2023). Furthermore, our data reveal traces of this correlation, though predominantly among the most successful bidders.

The idea is that when costs and quality are positively correlated, then by introducing a scoring auction, we put the firms closer to each other in the score dimension, thus lowering their market power (i.e., bid shading) and profits, as if more bidders have entered the market. However, a formal theoretical analysis suggests that mere correlation is not enough, as profits are fixed by the revenue equivalent principle as long as the ranking of firms remains unchanged across different auction formats. To change this ranking, we also require the firm's types to be bivariate. That is, firms must vary in two abilities: to reduce costs and to maintain quality standards.

Our explanation for the puzzle is straightforward. Although the rise in quality led to increased costs, this was offset by a reduction in bid shading, preventing an anticipated increase in price. However, since firms' profits have decreased, this adjustment was not a Pareto improvement, likely contributing to the political discontent and the eventual abandonment of the system.

Finally, we estimate the model non-parametrically to assess the magnitude of this effect in the data. Indeed, the model predicts that the switch to the scoring auction is associated with a decrease in the winning firm's bid shading by roughly 12 percent (€1.1 thousand per auction), on average, offsetting a significant portion of the increase in the winning firm's costs (€3.1 thousand per auction) due to selection. This means lower-than-expected prices for the buyer and lower-than-expected profits for the firms, which is in line with our explanation of the puzzle.

1.1 Literature

This study offers the first in-depth analysis of introducing such a vendor rating system in public procurement, i.e., scoring auctions with past performance.

Our paper also relates to the literature on industrial organization studies of auctions and competition. In particular, the need to use more complex auction systems relative to the standard price-only auctions is a key pillar of the literature on bidding for contracts. A vast theoretical literature has highlighted the limits of competitive auctions, starting at least from Spulber (1990) and including Manelli and Vincent (1995), Zheng (2001), Bajari and Tadelis (2001), D'Alpaos et al. (2013) and Burguet, Ganuza and Hauk (2012). This theoretical insight has found support in a handful of empirical studies, including Coviello, Guglielmo and Spagnolo (2018), Carril, Gonzalez-Lira and Walker (2020), Decarolis et al. (2020), and Bosio et al. (2020). Recently, new auction formats to trade off price vs. quality were proposed, see in Andreyanov, Krasikov and Suzdaltsev (2023) when quality is contractible and in Lopomo, Persico and Villa (2023) when it is not.

Several more empirical studies have also confirmed this result, highlighting how price competition can backfire in various contract performance measures ranging from quality to cost overruns and time delays.[5] Compared to this

[5]See Bajari, McMillan and Tadelis (2009), Decarolis (2014), Chong, Staropoli and Yvrande-Billon (2014), Liebman and Mahoney (2016), Lewis-Faupel et al. (2016), Kang and Miller (2021).

literature, our emphasis is on the use of past performance, which is novel.[6] Our findings are also related to a recent wave of studies highlighting the importance of considering dynamic incentives to understand procurement auctions. In this respect, this study is close in spirit to those of Jofre-Bonet and Pesendorfer (2003), and Chassang and Ortner (2016). Our theoretical model resonates with the seminal works by Che (1993), Asker and Cantillon (2008), and Asker and Cantillon (2010) on the design of scoring auctions, but with the focus on the investment costs.

The last strand of the literature is the design and use of contract audit measures. Detailed audit data on public procurement are used by Olken (2007) on Indonesia and Colonnelli and Prem (2021) on Brazil, as well as by Duflo et al. (2013, 2018) on environmental compliance. The mechanism that we study is based on third-party audits of past performance. Hence, it is also closer to the recent literature on the design of feedback mechanisms in platforms Tadelis (2016) than to the classic literature on reputation as an incentive to work hard to affect beliefs Klein and Leffler (1981); Holmstrom (1999). Still, our findings square well with the argument in List (2006) that reputation and quality verification are complements in that repeated interaction only increases the price/quality correlation when a quality rating system is present.

The paper is organized as follows. Section 2 provides an overview of the institutional details, the definitions of the reputation index, and the scoring formula. In Section 3, we show reduced-form evidence that quality has increased, but the price has decreased. We also perform a back-on-the-envelope calculation of the social value of the reform. In Section 4, we build a static model of a scoring auction that captures past performance as (sunk) investment costs to demonstrate the possibility of such price behavior. Finally, in Section 5, we estimate the model non-parametrically and simulate the reform to reinforce our findings. Section 6 concludes.

[6] A few theoretical studies have argued in favor of the positive role that reputation mechanisms linking the award of future contracts to the quality of past performance may improve contract performance in repeated public procurement under imperfect contracting. See, among others, Calzolari and Spagnolo (2009), Board (2011) and Andrews and Barron (2016).

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

Y2015	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/$)	3.2 (2.1)	4.4 (3.3)	4.9	3.1	3.0
Power Supply (TWh)	11	26	86*	29*	36*
Power Grid Works (mln/$)	206	318	2,400	500	475

*Note: 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). The values in parenthesis refer to power only for the total Turnover (bln $).*

2 Institutional Details

The buyer, Acea s.p.a., offers electricity and water services to about 1.6 million customers: private households and business establishments in the area of Rome. The firm is vertically integrated, owning and operating most 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). As shown in Table 1, all of these firms spend significant resources every year on works to preserve their power grid's operational efficiency.[7]

In 2015, Acea spent about US $200 million on procuring the kind of works that are the focus of this study. The jobs typically entail maintaining, upgrading, and replacing transformers, poles, underground cables, underground vaults, station transformers, and distribution and receiving stations. These are complex jobs requiring lots of manual labor and exposing workers to safety hazards linked to electricity-induced accidents. [8]

In 2007, after these risks materialized in some deadly accidents, Acea de-

[7]Acea and LADWP figures on employees and turnover include the water business. BGE and PECO figures on employees and turnover include the gas business.

[8]A search among local newspapers revealed that 4 workers had died in the last 15 years while performing works for Acea. The U.S. Bureau of Labor Statistics recorded 5,587 fatal electrical injuries between 1992-2013, an average of 254 fatal electrical injuries each year. Death was due to electrocution or fires caused by electricity, see Campbell and Dini (2015).

cided to take action to improve contract execution by revising its audit system. Until then, the auditors (i.e., a team of Acea engineers) inspecting the work sites used to prepare a written memo describing the state of the work site.

Notably, the reform only involved the electricity sector, leaving out the water sector.

Audits.

On October 16, 2007, Acea's engineers conducted their first surprise audit under the vendor rating system, which streamlined and digitized the process: using tablet computers, the inspection required evaluating a fixed list of 136 parameters by scoring them as pass, fail, or uninspectable.

The list of 136 parameters, organized into 12 categories, was identified as exhaustive of the quality and safety standards that needed to be audited: they ranged from the types of materials and machinery used to the adherence to the worksite safety specifications and legal status of all workers (the full list is reported in Appendix F, and the 12 categories also in Table 2). The logic followed by Acea was to cover with these 136 parameters all of the relevant features of contract performance.

Randomization was implemented at two levels to limit the risk of corruption and biased evaluations. First, the work sites to inspect every week were randomly drawn from those where suppliers were actively working. Thus, the same worksite could be audited multiple times or never. Second, the composition of the 3-member auditor teams was randomly drawn from the pool of Acea auditors (around 12 engineers).

Penalties were formally always part of the contracts, both before and after the reform of the audit system. However, they were rarely enforced to avoid taking disputes to the civil court of law, as conveyed by the Acea personnel.[9]

Announcements.

Another peculiar element of the reform's timing is that Acea initially con-

[9] According to the National Statistical Institute (ISTAT), in the period between 2002 and 2012, the length of this type of proceedings in front of the Rome court was, on average, 3 years and 3 months. Two levels of appeal can make it even longer. See Coviello et al. (2018) for the discussion of the effects of court delays on contract penalties enforcement.

cealed its motivation for switching to digitized audits. It was only three months after the new audit system had started that Acea announced to its suppliers in a public meeting held on December 20, 2007 ($t1$) the intention to switch its contract procurement system from price-only auctions to scoring auctions. This allowed Acea to evaluate the distribution of RI in the price-only auctions.

In five consecutive public meetings with the suppliers ($t1, t2, \ldots, t5$), Acea explained this new system and showed simulations of how a firm would benefit from higher RI. It also privately informed each firm of its current RI and the distribution of RI across all firms.

Reputation index.

In each audit, auditors evaluate all feasible parameters and the condition of the worksite: on average, 34 parameters are evaluated per auditor visit. Each evaluated parameter p is given a value of 0 or 1. We will denote the corresponding value as v_{ap}. Each parameter is assigned a weight w_p between 1 and 10 when it is being evaluated and 0 otherwise, in which case it is absent from the data. These weights are constant across audits.

For any given firm, its reputation index (RI) at the moment t is a rolling average of its weighted evaluations

$$\mathrm{RI}_t = \frac{1}{m} \sum_{a=1}^{m} \frac{\sum_{p=1}^{136} v_{ap} w_p}{\sum_{p=1}^{136} w_p}. \tag{1}$$

The audits $a = 1, \ldots m$ considered for the calculation of RI are those in the 12 months before t.[10]

Scoring formula.

The RI is part of the following scoring formula[11], which determines the

[10]For new entrants and firms with very sparse audit data, Acea decided to calculate the RI only if at least seven audit visits had been done in the previous 12 months; otherwise, the supplier would be assigned an RI equal to the average RI of the bidders in the auction. 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. This rule limited a "cold start" problem as a barrier to entry, as discussed in Butler et al. (2020).

[11]This scoring rule is equivalent to: Score = Discount + $\alpha \cdot RI$, where $\alpha = \frac{\omega}{1-\omega}$.

winner of the scoring auction as the bidder with the highest score:

$$\text{Score} = (1-\omega)\cdot \text{Discount} + \omega \cdot \text{RI} \tag{2}$$

$$\text{Discount} = 1 - \frac{\text{Price offered}}{\text{Reserve price}}, \quad \omega = 0.25. \tag{3}$$

A price-only auction is a scoring auction with weight $\omega = 0$.

In total, 36 scoring auctions were held in roughly three months, with an average participation of 13.6 bidders.

3 Descriptive and reduced-form evidence

The analysis is based on several sets of data: Acea audit and auction data, Arera and Istat external measures data, and Telemat auction data.

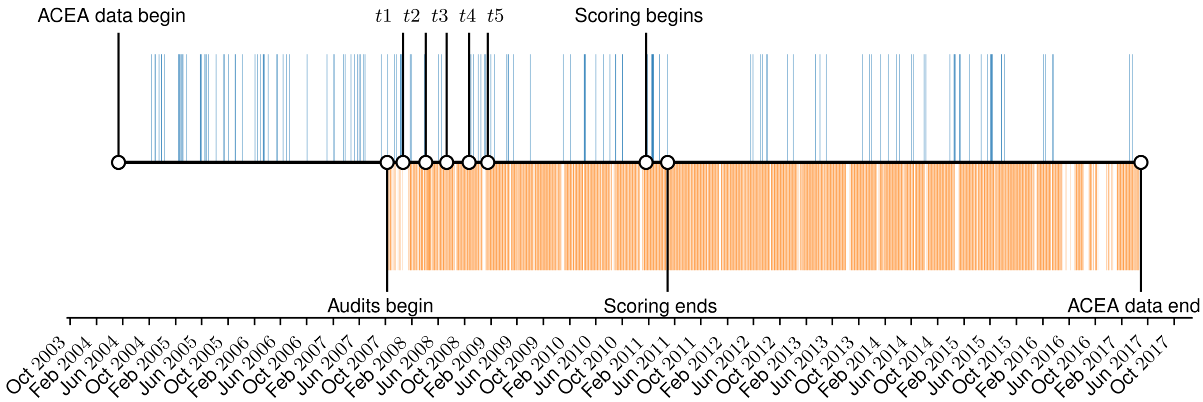
The first two datasets come from Acea, the vendor rating system designer, containing detailed audit data recorded through the old and new auditing systems and detailed auction data, including RI and bids of all participants. The following two datasets come from Arera and Istat - public authorities supervising the power and water sectors and contain external performance measures. The last dataset comes from Telemat, a large provider of public tender data, containing auction data covering bidding and other auction-related information. The Telemat auction data is less detailed than Acea's, as companies in Italy are only legally bound to report the winning bid, but not the losing bid or the total number of bids.

The timeline is split into four main periods, see Figure 1: price-only - before December 2007 (t1), grace - between December 2007 (t1) and January 2011 (scoring begins), scoring - between January 2011 (scoring begins) and May 2011 (scoring ends), and hybrid - after May 2011 (scoring ends).

Internal performance measures.

Acea's audit records span from 2007 to 2017, with 365,896 values assigned during 8,973 audits involving 634 contracts and 84 different contractors. Since worksites are randomly inspected weekly, a contract might receive no inspections or multiple inspections over its life.

Figure 1: Timeline



Note: upward blue ticks represent auctions; downward orange ticks represent audits. Acea's five announcements of the future switch to equation (2) are marked as $t1$,$t2$,$t3$,$t4$,$t5$.

This data allows us to observe the evolution of quality captured by the internal performance measures before, during, and after the scoring period. Table 2 contains average parameter compliance across 12 main categories and 4 time periods (1 corresponds to maximal full compliance). It is also useful to follow the evolution of quality aggregated into a single metric, such as the RI. Since we do not have RI records outside of the scoring period, we reconstruct it, as close as possible, following the definition of RI and the available data.

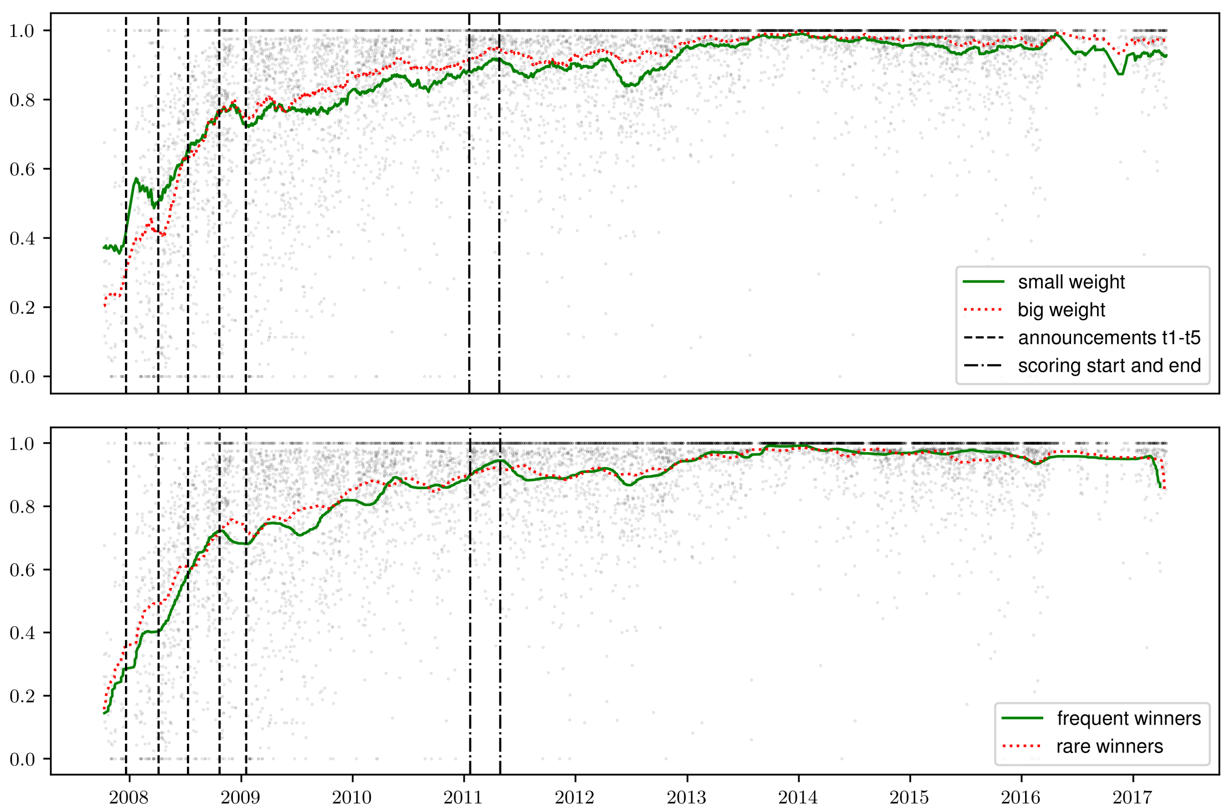
In the first exercise, parameters were split into two categories: small and large weights w_p. The analog of RI was reconstructed separately for each group. Figure 2 (top) shows that compliance with more important parameters (large weights) was initially lower than with less important ones (small weights), likely due to differing perceptions of performance between firms and Acea. Over time, firms realigned their behavior, focusing more on large-weight parameters.

Table 2: Internal Performance Measures by Period and Category.

Category	Price-only	Grace	Scoring	Hybrid
Air works		.98	1	1
Customer relationship mgnt	1	.93	1	1
Documentation	.35	.66	.83	.93
Equipment and machinery	.7	.93	.96	.95
H.T. works site controls		.8	.96	.96
Personnel	.32	.69	.88	.96
Underground works	.4	.69	.85	.81
Works execution	.19	.84	.96	.98
Works on joints	1	.96	.94	.86
Works on transformer station	1	1	1	1
Works site regularity	.11	.62	.83	.94
Works site safety	.31	.76	.91	.96

Note: the table shows the share of compliant parameters (1 corresponds to full compliance).

Figure 2: Internal Performance Measures by Weight and Firm Type.



Note: the plots represent the moving average (centered, with a 60-day window) of the audit results, weighted (within a respective group of audits) proportionally to the definition of RI.

Table 3: Summary Statistics for the External Performance Measures.

	Mean	St. Dev.	Median	Max	N
LLB (num/LV lines)	2.43	2.50	1.76	24.00	1,433
LLB duration (min)	94.03	134.01	49.40	960.00	1,419
SLB (num/LV lines)	2.70	3.90	1.84	62.00	1,286
PPC (num/LV lines)	0.60	1.24	0.30	29.47	1,431
PPC duration (min)	65.63	113.96	31.18	988.53	1,428
Low voltage users (000)	365.45	815.30	6.42	4663.64	1,642
Water Leakage (%)	0.33	0.09	0.32	0.59	253
Water users (000)	899.80	1061.58	491.33	4341.24	253

Note: external performance measures recorded by Arera between years 2000 and 2016.

In the second exercise, firms were divided into frequent winners in price-only auctions (top 10 before 20 December 2007) versus others. Figure 2 (bottom) shows that all suppliers improved performance, but the frequent winners in price-only auctions took longer to comply. This indicates that firms that are better at cost reduction are not necessarily better at producing quality, reflecting quality concerns (adverse selection) in the market.

Acea's audits show that the suppliers actively responded to the new incentives provided by the vendor rating system.

External performance measures.

In Italy, both the electricity and water sectors are subject to partial regulation. Regulation primarily persists in the transmission domain of electricity; however, the regulatory authority Arera gathers comprehensive data across the entire sector. Through Arera, we obtained various yearly firm-level performance metrics, encompassing all low-voltage power distributors.

Key performance metrics include the frequency and duration of long-lasting blackouts (LLB), short-lasting blackouts (SLB), and programmed power cuts (PPC).[12] These metrics are presented in the upper five rows of Table 3. Notably, these metrics do not overlap with any of the RI parameters.

[12]LLB and LLB duration are, respectively, the average number and the average duration (in minutes) of long-lasting blackouts per user, SLB is the average number of short-lasting blackouts per user, PPC and PPC duration 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), *Water Leakage* 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).

Table 4: Estimates for the External Performance Measures

	(1) LLB	(2) LLB duration	(3) SLB	(4) PPC	(5) PPC duration	(6) Water leakage	(7) Water leakage
β_0	-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)
Obs.	386	386	298	386	386	253	59
R^2	0.84	0.57	0.83	0.72	0.79	0.82	0.89
Buyer and Year FE	✓	✓	✓	✓	✓	✓	✓
Sample	All	All	All	All	All	All	Pop $\geq$ 1mln

Note: estimates for regression (4). Robust standard errors in parentheses.

We assess the impact of Acea's 2007 announcement on the set of external performance metrics using a difference-in-differences (DID) approach, as delineated by the following model:

$$O_{ft} = a_f + b_t + \beta_0 D^{Acea}_{t \geq t_1} + \gamma X_{ft} + \epsilon_{ft}, \tag{4}$$

where O_{ft} represents observed performance outcomes for each distributor, f, in each year, t. Here, a_f and b_t stand for fixed effects for distributors and years respectively, X_{ft} includes controls such as customer number, and $D^{Acea}_{t \geq t_1}$ is a dummy variable indicating Acea's auctions post-2007.[13] The coefficient of interest, β_0, captures the differential performance of Acea post-announcement, relative to other low-voltage distributors.

Table 4 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 cover the water distribution sector. One can see that planned power cuts became longer, likely due to increased compliance to safety and quality, while the unplanned power

[13] 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 all the distributors (column 6) or only those in charge of geographical areas with at least 1 million customers (column 7).

cuts (i.e., blackouts) became shorter and less frequent. At the same time, the water sector, unaffected by the reform, shows no improvements.

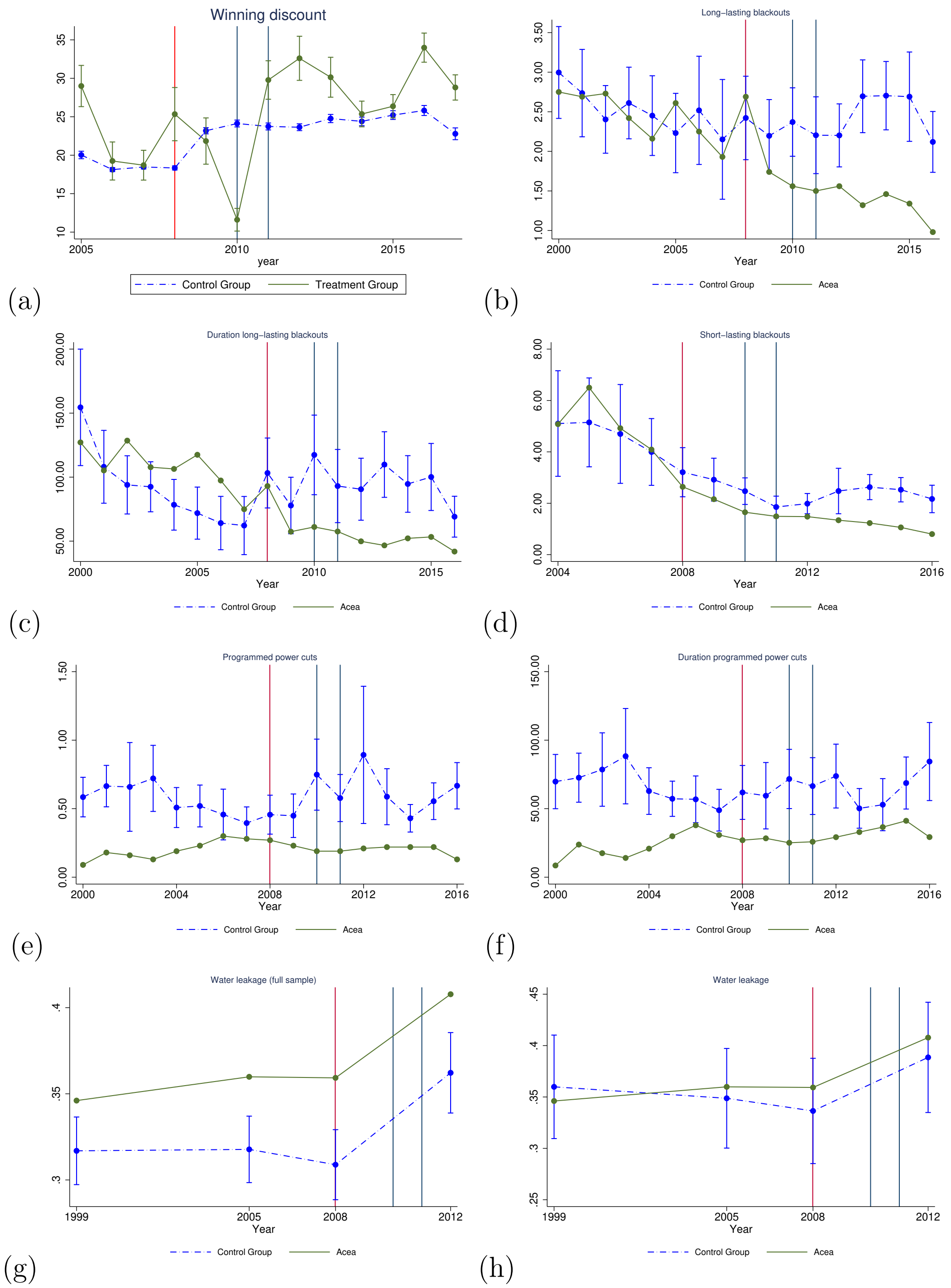
Furthermore, Figure 3 depicts, through various panels (b) to (f), the trends across our identified external measures, such as the frequency and duration of both long-lasting (exceeding 3 hours) and short-lasting (below 3 hours) blackouts, and scheduled power interruptions. These trends show Acea's relative improvement post-target year ($t1$), exemplified by a decrease in the number of long-lasting blackouts, both in absolute and relative terms compared to the control group. In panel (a), we also observe the evolution of winning discounts, which will be studied later in greater detail.

Scheduled power cuts effectively enhance service quality by substituting unplanned outages with planned maintenance. The data supports the inference that improvements by Acea's suppliers in external performance metrics could follow enhancements in performance, albeit with potential delays due to technological constraints. For instance, upgrades in grid materials might only manifest in reduced outage frequencies after substantial portions of the network are upgraded.

The latter panels of Figure 3, specifically (g) and (h), direct attention to the water sector, revealing trends in water leakage rates, which have remained outside the audit reform. This sector thus acts as a placebo in our analysis. Data from the Italian Statistical Institute (Istat) environmental census, which records water inflow and outflow across Italian counties between 1999 and 2012, allows the measurement of water leakages as a percentage of total water inflow.[14] The bottom rows in Table 3 report summary statistics for the water sector, while the bottom panels of Figure 3 plot the dynamic of the water leakage indicator, separately for Acea and other firms indicating that there is no evidence of lower leakages for Acea.

[14]Given legal constraints that restrict each county to a single water distributor, aggregation of county-level data to reflect Acea's footprint is straightforward. 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.

Figure 3: Evolution of Discounts and External Performance Measures



Note: Acea (in green) and other utilities (in blue, dashed). The red line indicates the date of Acea's first announcement. The blue, vertical lines indicate the scoring rule period.

Overall, the external measures confirm the long-lasting performance improvements associated with the reform.

Social value of the reform.

To grasp the scale of the value of the reform, we conduct a back-of-the-envelope calculation focusing on the frequency of fatal accidents and blackouts.

We start with the duration of long-lasting blackouts. We convert the estimate in column 2 of Table 4 into a measure of blackouts 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),[15] we associate a cost of blackouts of 2.5 euro/hour for residential customers and 18.75 euro/hour for business customers. The result is that the reduction in blackouts implies a benefit of 6.6 million euros/year.

Next, we focus on the change in the probability of fatal accidents as implied by improvements in the safety related parameters. To map the changes in safety compliance to the occurrence of fatal accidents, we use the statistical model employed by Acea's engineers — the *Heinrich's Pyramid*.[16] It estimates the reduction in the probability of a fatal accident as 0.54-0.82 per year. With 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, the estimated safety benefits range between 3.5 and 5.3 million euro/year.[17]

Thus, our estimates place the lower bound on the social value of the reform between 10.1 and 11.9 million euro/year.[18]

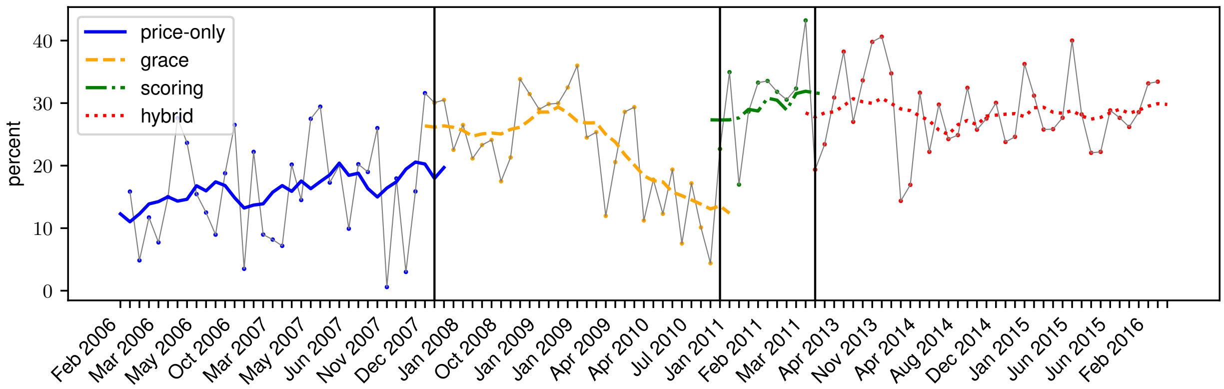
[15]See Arera's decision n. 172/07 of 12/07/2007.

[16]See Appendix B for the definition of the *Heinrich's Pyramid*

[17]This subset of safety-related parameters was decided by Acea's engineers and is identified with an * in Table A.4 in Appendix F. The number of workers present on the worksite was estimated for us by expert engineers. The OECD (2012) values are converted to the 2007 nominal euro.

[18]This range is not our interval estimate, but the result of using the two bounds in the definition of Heinrich's pyramid

Figure 4: Evolution of Discounts Over Time.



Note: average (per auction day) discount of Acea's winning firm and it's moving average.

Acea's auctions.

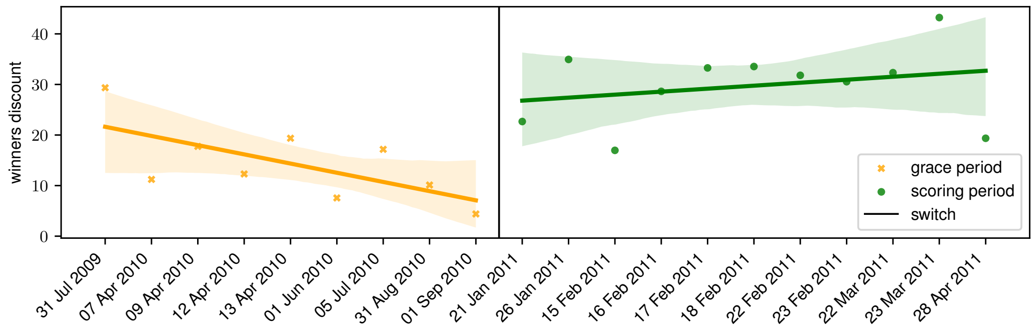
The records of Acea's auctions span between the years 2004 and 2017. We are primarily interested in the evolution of Acea's winner's discounts. Unlike both internal and external performance measures, which increase steadily and monotonically over time, the discounts change non-monotonically. To obtain a smooth trajectory of discounts, we use a moving average (window centered, over five auction days), see Figure 4.[19]. The cumulative effect appears positive if we compare the start of the grace period to the end of the scoring period.

If we focus on the grace period, the discounts grow and then rebound. We believe the initial increase is due to the temporary stimulus to win more auctions. Firms need to win more auctions to increase their chances of being audited and increase one's reputation index. Indeed, during this time, we observe the largest increase in RI. However, closer to the end of the grace period, the discounts fall back to levels similar to or even below those observed before the first announcement (t1).

An important detail is an apparent discontinuity around the switch from grace to the scoring period. It is observed for both average and winning discounts but is higher for the latter. Unfortunately, the auctions are located

[19] Moving average corresponds to rectangular- kernel smoothing. Boundary correction (via reflection) is applied to express the potential discontinuities between the periods.

Figure 5: Evolution of Discounts Around the Switch.



Note: average (per auction day) discount of Acea's winning firm and the 95% CI.

sparsely over time, preventing us from formally testing continuity. Still, we can estimate two regression lines in the small neighborhood of the switch, see Figure 5.[20] Indeed, we observe a sharp discontinuity. If we average over three auction days before and after the switch, the discount jumps up by approximately 14 percentage points (from 11 to 25).

The jump indicates that the switch in the auction mechanism is likely behind the rise in discounts. However, several exogenous factors could also contribute to the discount: (i) the reserve price, since the discount is the percentage distance between the winning bid and the reserve price; (ii) the number of bidders, since the winning bid is the lowest order statistic of bids.

It is unlikely that the reserve price explains the oscillation of the discount. First, 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 yearly by the region where Acea operates and used exclusively by contracting authorities to calculate reserve prices, thus excluding deliberate manipulation. These prices are much higher than the typical production costs, so the reserve price is always non-binding.

On the other hand, the number of bidders contributes to the winning discount in two different ways. First, as it is an order statistic, the winning

[20] The running variable is the order of the auction day.

Table 5: Relationship Between Offered Discount and Reputation Index

	(1)	(2)	(3)	(4)
β_1	0.135***	0.111***	-0.060***	-0.061***
	(0.015)	(0.024)	(0.019)	(0.021)
Obs.	1699	1699	1699	1699
R^2	0.047	0.420	0.862	0.903
Bidder Rank		✓	✓	✓
Auction FE			✓	✓
Lot FE				✓

Note: estimates for regression (5). Robust standard errors in parentheses.

discount is increased by the number of bidders. Second, if the participants observe or anticipate an increase in bidders, they bid more aggressively, increasing the discount even further. Although our data shows no noticeable jump in the number of bidders in the neighborhood of the switch, we repeat the exercise with residualized discounts, see Appendix E.

Another possible explanation for the change in discounts is the change in the composition of the pool of participants in auctions. However, because Acea is a public buyer and because of the EU and Italian Regulations, the entry barriers into these auctions are small, so endogenous entry is negligible.

Before moving to the next section, we ask one last question. Is the discount related positively or negatively to the reputation index? To answer this question, consider the equation below:

$$D^o_{ijk} = a_j + b_k + \beta_1 RI + \beta_2 RI \cdot Rank + \beta_3 Rank, \tag{5}$$

where D^o_i is the offered discount, the index i indicates the bid, j is the auction and k is the lot. Fixed effects are the auction (a_j) and lot (b_k). The sample are the price-only auctions held during years 2007-2016, with RI reconstructed using audit data.[21] *Rank* is a numerical variable representing the rank of the bidder in the price-only auction (rank 1 corresponds to the winner, rank 2 corresponds to second best, etc.). We are primarily interested in β_1.

[21] By matching the audits to auctions, we were able to assign an RI to 62% of bids in the price-only and grace periods.

Table 6: Winning Discount by Period

	Price-only	Grace	Scoring	Hybrid
Acea	21.93 (10.41)	18.68 (10.48)	28.84 (7.20)	28.65 (7.8)
Control	18.52 (9.6)	21.07 (11.97)	24.75 (13.27)	24.45 (13.53)

Note: the table contains average winning discount and standard deviations in parentheses.

The estimation results are presented in Table 5. If we pool all bidders (winners and losers) together, which corresponds to specification (1), the regression coefficient of the offered discount on the reputation index is positive and equal to approximately 0.13. However, if we consider the rank of the bidder in the price-only auction, which corresponds to specifications (2)-(4), it turns out that, among the most competitive firms, the correlation is negative and equal to -0.06.[22]

Since the strongest participants determine the outcome of the auction, we interpret this as strong evidence of quality concerns, similar to Lopomo, Persico and Villa (2023), in the environment.

Telemat auction data.

To further assess the change in Acea's prices, we look at similar contracts[23] by other companies, which we refer to as a control group. This gives us the opportunity to account for general trends that might explain some of the price movements.

The Telemat data include the object of the contract, the reserve price, the winning discount 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.[24]

One can see from Table 6 that the discounts indeed exhibit a positive trend

[22] For comparison, the slope of the scoring rule would correspond to -0.33.

[23] These contracts belong to a well-defined category identified by the Italian regulation as "$OG10$," which makes it feasible to select comparable projects across different buyers. Furthermore, using textual search methods, we separated $OG10$ 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., those with a reserve price below €10,000 or above €2.5 million).

[24] The number of bidders is missing in Telemat data and so is imputed with a constant.

Table 7: Treatment Effect of Scoring Using Telemat Data

	(1)	(2)	(3)	(4)
β_6	4.15	4.01	5.17	4.69*
	(3.15)	(3.09)	(3.68)	(2.61)
Obs.	41482	13505	2608	11294
R^2	0.54	0.63	0.58	0.64
Buyer&Year FE	✓	✓	✓	✓
Size, region, category		✓	✓	✓
Control sample	All	All	Center	North&South

Note: estimates for regression (6). The sample includes auctions by Acea (treatment group) and all other contracting authorities (control group). Standard error clusters by year and CA are reported in parentheses.

in the control group, which could contribute negatively to Acea's discounts in the scoring period. Thus, it makes sense to estimate the treatment effect of the scoring period in Acea's auctions relative to the control group.

Consider the equation below:

$$D_{ift}^{w} = a_f + b_t + \beta_4 d_{t1-t5}^{Acea} + \beta_5 d_{grace,>t5}^{Acea} + \beta_6 d_{score}^{Acea} + \beta_7 d_{hybr}^{Acea} + \gamma X_{ift} + \epsilon_{ift}, \quad (6)$$

where D_{ift}^{w} is the winning discount, the index i indicates the auction, f is the entity awarding the contract, and t is the year. There are four treatment dummy variables for the contracts awarded by Acea in each of the 4 periods: grace before t5, grace after t5, scoring, and hybrid. The base group is, therefore, the price-only period before the announcements. Fixed effects are the entity awarding the contract (a_f) and year (b_t) and other controls (X_{ift}).[25] We are primarily interested in β_6.

The estimates are presented in Table 7 in four specifications differing in the set of covariates and control group observations. In particular, we consider limiting the sample to either buyer located in central Italy (which might be more similar to Acea in terms of input prices, the pool of suppliers, and environmental conditions) or only outside this area (which might serve to limit

[25] Additional controls involve contract characteristics, a dummy for whether the reserve price surpasses the thresholds: 250000, 500000, 1500000 or 5000000; a dummy for whether the contract is for public illumination, and the logarithm of the number of bidders + 1.

contamination concerns).

Across all specifications and samples, we find fairly consistent estimates: Acea's discounts are still greater in the scoring period than in the price-only period, albeit not always statistically significant.

4 Stylized model

In this section, we propose a stylized private-information model of a scoring auction, where non-price characteristics of the firm are related to its past performance. The firm can invest in its quality by performing better. However, since performance is measured in past contracts, the associated costs are effectively sunk from the auction viewpoint. This differs from the classical models of scoring auctions in Che (1993) and Asker and Cantillon (2010), where costs of raising quality are not sunk. The model will be ex-ante symmetric, allowing us to make sharp predictions using a unique symmetric Bayes-Nash equilibrium.

Let the auction have n ex-ante identical firms competing for a single procurement contract. The reserve price is normalized to 1 and is non-binding[26]. Let $0 \leqslant \theta_i \leqslant 1$ be firm i's cost-efficiency parameter, and $q_i \geqslant 0$ be her observed quality (i.e., her past performance). Furthermore, let the firm have convex production costs $C^P(\theta, q)$ that are strictly decreasing in θ_i but strictly increasing in q_i, and convex investment costs $C^I(\underline{q}, q)$. The $(\theta_i, \underline{q}_i)$ pair captures the firm's private type and is i.i.d.

We will consider a *quasi-linear scoring rule* $s_i = \alpha q_i + d_i$, where s_i is the firm's score and d_i is its discount (i.e., the difference between the reserve price and its bid b_i).[27] As in all scoring auctions, the firm with the highest score wins the contract.

We assume that there is no exchange of information between the firms after they invest in quality and before they choose their discounts. Thus

[26] For simplicity, we allow for negative discounts.

[27] Quality q corresponds to the reputational index RI in the data, and the α weight (often referred to as the *the dollar value of quality*) is related to the ω weight as $\alpha = \frac{\omega}{1-\omega}$.

the choice of (q_i, d_i) or, equivalently, (q_i, s_i) can be modeled as simultaneous. Following Asker and Cantillon (2010), we will create an auxiliary variable called the *pseudo-type* $\rho_i = \alpha q_i + 1 - C^P(\theta_i, q_i)$, such that firm $i's$ profit margin $b_i - C^P(\theta_i, q_i)$, upon winning the auction, is equal to $\rho_i - s_i$.

We are interested in a symmetric equilibrium with strictly monotone strategies $\sigma : \rho \to s$. Denoting the equilibrium distribution of score as $F_s(.)$, then each firm maximizes

$$U_i(q, s) = (\rho_i - s)G(s) - C^I(\underline{q}_i, q), \quad G(s) = F_s^{n-1}(s)$$

subject to her type $\theta_i, \underline{q}_i$ and the $\rho_i = \alpha q + 1 - C^P(\theta_i, q)$ constraint.

This implies two sets of first-order conditions:

$$(\rho - s)\frac{\partial G}{\partial s}(s) = 0, \tag{7}$$

$$\alpha G(s) - \frac{\partial C^I}{\partial q}(\underline{q}_i, q) = 0. \tag{8}$$

Equation (7) is the standard optimality condition for auctions. It also shows that the score depends only on the pseudo-type ρ since there is no binding reserve price. The pseudo-type is, however, endogenous. Still, the equilibrium strategy can be written as

$$\sigma(\rho) = \int_{\underline{\rho}}^{\rho} z dH(z)/H(\rho), \quad H(\rho) = F_\rho^{n-1}(\rho) \tag{9}$$

where $F_\rho(.)$ is the (endogenous) equilibrium distribution of pseudo-type and $\underline{\rho}$ is the lowest participating pseudo-type. Equation (8) is the condition for the optimal choice of quality, which is both necessary and sufficient since U_i is strictly concave in q, for any score s.[28] We will denote the solution to $\frac{\partial C^I}{\partial q}(\underline{q}_i, q) = \alpha h$ as $q^*_\alpha(\underline{q}, h)$, which is monotone in h i.e., the conjectured probability of winning.

Our first observation is that there must be full participation. Indeed, if

[28] Contrary to Che (1993), where the choice of quality was independent of the score, here the marginal cost of quality equals α (the dollar value of quality) times $G(s_i)$.

a positive mass of types does not enter in a symmetric equilibrium, then the auction has no participants with a positive probability. Moreover, the reserve price is, by construction, not binding. Thus, any potential entrant can profitably enter with a large enough bid.

Consider the function $\rho^*_\alpha(\underline{q}, \theta) := \alpha q^*_\alpha(\underline{q}, \theta^{n-1}) + 1 - C^P(\theta_i, q^*_\alpha(\underline{q}, \theta^{n-1}))$ which uniquely identifies the pseudo-type ρ as a function of type θ. One can verify whether this function is monotone in θ.[29] With the distribution of pseudo-type at hand, we can compute equilibrium quality and score via equations (8) and (9), which completes the derivation of the equilibrium. It remains to show that the second-order conditions are satisfied, see Appendix C.

Proposition 1. *If $\rho^*_\alpha(\underline{q}, \theta)$ is monotone in θ, for all $(\underline{q}, \theta)$ in the support, a symmetric equilibrium with a strictly monotonic strategy σ exists and it is unique. Moreover, there is full participation of types.*

We split the remaining analysis into two separate cases.

Univariate types.

For now, let the firms vary only in their cost-efficiency parameter θ.

Recall that the score is monotone in pseudo-type, which is, in turn, monotone in type, conditional on $\underline{q}$. Thus, the firm with the highest θ wins, independently of α. This leads to several important conclusions considering an anticipated switch from a price-only to a scoring auction.

First, the quality of every firm will go up by (8). Consequently, the winner's expected quality has to increase. On the other hand, the firm's interim expected utility is fixed by the revenue equivalence principle. Thus, her discount has to decrease to compensate for the increased investment costs. Consequently, the winner's expected discount has to decrease.

See the example below for an illustration.

Example 1. *Let there be 2 firms,*

$$\underline{q} = 0, \quad C^P(\theta, q) = 1 - \theta, \quad C^I(\underline{q}, q) = (q - \underline{q})^2/2,$$

[29] In particular, it is true if $C^P(\theta_i, q) = 1 - \theta$.

and θ distributed uniformly on $[0, 1]$.

Assuming monotonicity of score in type, $G(s(\theta)) = \theta$, we can compute the equilibrium quality $q = \alpha\theta$ and the equilibrium pseudo-type $\rho = (1 + \alpha^2)\theta$. On the other hand, the total equilibrium profit of the firm is equal to $\int_0^\theta x dx = \theta^2/2$ by the envelope conditions. It does not depend on α because the firm with the highest type always wins. The auction profits are, therefore, equal to the total profits $\frac{\theta^2}{2}$ plus the investment costs $\frac{(\alpha\theta)^2}{2}$. Dividing the auction profits by the probability of winning, we can compute the profit margins $(\rho - s) = (1+\alpha^2)\theta/2$. The equilibrium score is, therefore, $s = (1 + \alpha^2)\theta/2$, and the discount is

$$d = (1 - \alpha^2)\theta/2,$$

which is decreasing in α.

Note that in the example above, current production costs are independent of q. This is not an unreasonable assumption since we interpret q as past performance. More generally, we can allow the production costs to increase in q, capturing the inertia between past and current performance. Moreover, we can also allow for a functional relationship between θ and $\underline{q}$, as long as the pseudo-type ρ is uniquely defined by and is monotone in type θ.

Example 2. *Let there be 2 firms,*

$$\underline{q} = 1 - \theta, \quad C^P(\theta, q) = 1 - \theta + q, \quad C^I(\underline{q}, q) = (q - \underline{q})^2/2,$$

and θ distributed uniformly on $[0, 1]$.

Assuming monotonicity of score in type, $G(s(\theta)) = \theta$, we can compute the equilibrium quality $q = (1 - \theta) + \alpha\theta$ and the equilibrium pseudo-type $\rho = (1 + (\alpha - 1)^2)\theta + (\alpha - 1)$. The auction profits are, again, equal to the total profits $\frac{\theta^2}{2}$ plus the investment costs $\frac{(\alpha\theta)^2}{2}$. Dividing the auction profits by the probability of winning, we can compute the profit margins $(\rho - s) = (1+\alpha^2)\theta/2$. The equilibrium score is, therefore, $s = \frac{(a-1)(2-3\theta+\theta\alpha)}{2}$, and the discount is

$$d = \frac{-2 + 3\theta - 2\theta\alpha - \theta\alpha^2}{2},$$

which is decreasing in α.

The change in discounts has two main components here. The first is the compensation for the investment costs associated with acquiring higher quality prior to the auction. The second is the increase in production costs associated with executing the contract given higher quality. Thus, the combination of the two produces an unambiguous decrease in discounts.

Proposition 2. *With univariate types, in the scoring auction, the expected quality is higher, and the expected discounts are lower than in the corresponding price-only auction.*

Thus, univariate types can not explain our empirical puzzle.

Bivariate types.

Let the firms vary in both θ_i and $\underline{q}_i$, but there is no functional relationship between the two variables. Relative to univariate types, this introduces two new sources of variance: the selection of winning firms and bid shading.

Consider an anticipated switch from the price-only auction to the scoring auction. Since the former is cost-efficient, the winning firm's θ_i can only decrease, and strictly so for generic distributions. Thus, if there were no strategic adjustment of bids or quality, an increase in α would guarantee an increase in costs and, therefore, a decrease in discounts, just like with univariate types.

However, the strategies (9) can go either way. If, with the introduction of the scoring auction, the pseudo-type distribution becomes more concentrated, the shading is likely to decrease due to a more aggressive equilibrium strategy σ. Thus, while a switch to the scoring auction necessarily leads to an increase in the production costs (at every level of q) of the winning firm, it may lead to an increase in discounts if the shading of the winning firm decreases significantly. This can be achieved when q and θ are slightly negatively correlated.[30]

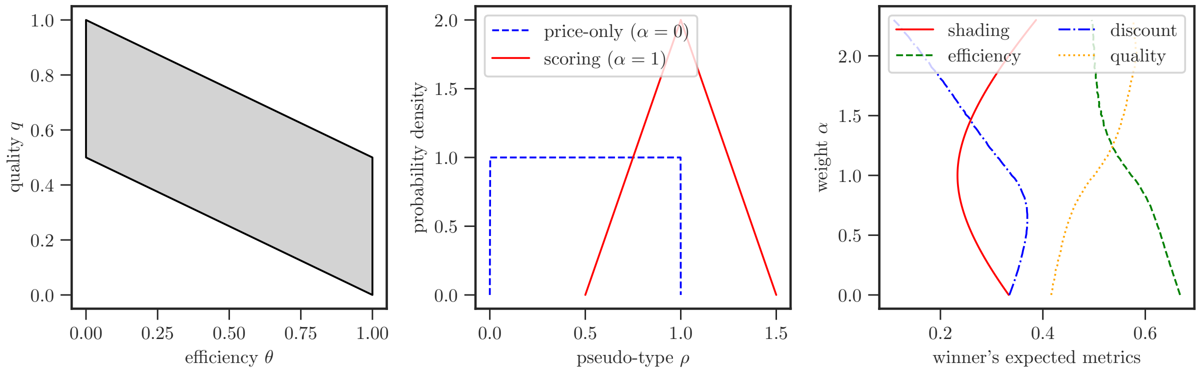
To illustrate the idea, see the example below

Example 3. *Let there be 2 firms,*

$$C^P(\theta, q) = 1 - \theta, \quad C^I(\underline{q}, q) = (q - \underline{q})^2/(2\beta)$$

[30]Negative correlation per se is not sufficient to explain the puzzle, see Example 2.

Figure 6: Bivariate Types Example



and a uniform distribution of (θ, q) *in the scoring auction, in the region defined by* $0 \leqslant \theta \leqslant 1$ *and* $1/2 \leqslant q + \theta/2 \leqslant 1$*, see Figure 6 (left).*[31]

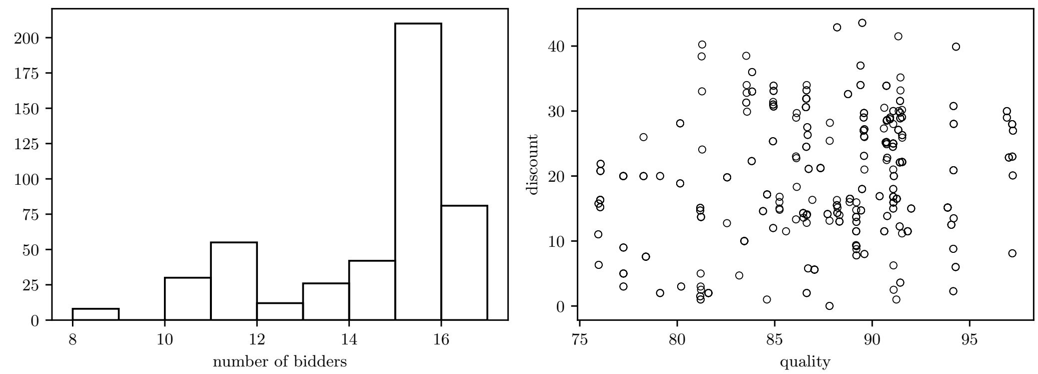
For the price-only auction, that is, a scoring auction with weight $\alpha = 0$*, the pseudo-type distribution is uniform, while with weight* $\alpha = 1$*, it is pyramid shaped, see Figure 6 (middle). The latter is more concentrated and is therefore associated with more aggressive bidding (i.e., less shading). While the expected winner's cost has changed from 20/60 to 23/60, her shading has changed from 20/60 to 14/60. As a result, the winner's discount has increased from 20/60 to 23/60, see Appendix C for the derivation.*

We also simulate numerically the evolution of the winner's characteristics in the example above, as functions of α, see Figure 6 (right), holding firm's quality fixed (which corresponds to β approaching 0). Indeed, for a range of weights between 0 and roughly .6., both expected quality and expected discount are increasing.

Thus, the bivariate types model can explain our empirical puzzle.

[31] To rationalize the observed distribution of (θ, q), one has to verify that $\rho - \alpha^2 \beta H(\rho)$ is increasing for all ρ in the support. For $\alpha = 1$ it would suffice that $\beta < 1/2$. For simplicity, one can assume that quality is exogenous, which corresponds to $\beta \to 0$.

Figure 7: Distribution of Auction Data



5 Empirical model

In this section, we estimate a structural model of a scoring auction non-parametrically and simulate the outcomes of a counterfactual price-only auction. We are primarily interested in the behavior of discounts across the counterfactual simulations.

The dataset consists of 34 first-score sealed-bid auctions held over 11 days between 2011-01-21 and 2011-04-28. The scoring rule is quasi-linear with weight $\alpha = 1/3$. We observe 479 quality-discount pairs, with quality (measured by the reputational index RI) distributed above 76 (out of 100) while discount distributed below 44 (out of 100). The number of bidders varies between 8 and 16, with an average of 13.64 and the mode at 15, see Figure 7.

To pick an appropriate structural model, we have to answer four key questions: (i) whether the reserve price is binding, (ii) whether the number of bidders is known, (iii) which model of heterogeneity to use, and (iv) whether this is an IPV (independent private values) or APV (affiliated/correlated private values) environment.

The environment suggests that the answer to the first two questions is negative. Indeed, the reserve price is intentionally set so that it is almost never binding. Moreover, since the format is sealed-bid, firms do not have hard information about who participates, so it makes sense to model the number of bidders as random. This also allows us to nearly double the sample size.

Next, using a standard mapping[32] between the first-price and the first-score auctions when the reserve price is not binding, we can use the tests from Krasnokutskaya (2011) to pick a suitable model of auction-level heterogeneity. The additive model of heterogeneity is soundly not rejected, see Appendix D. The intuition behind the additive model is that each contract has a fixed cost common to all bidders in the auction. The variability in the scale of production costs is of lesser concern since discounts are already measured as a percentage of the reserve price.

Finally, we would like to test whether, conditional on the observables, this is an IPV rather than an APV model. We apply the analog of the sup-norm test, suggested by Haile, Hong and Shum (2003), and the IPV hypothesis is not rejected, see Appendix D.

Model primitives and identification

Consider a single representative auction, as if there is no auction-level heterogeneity and quality is already chosen and observed by the buyer. As in the stylized model, denote quality as q, discount as d, pseudo-type as ρ, score as s, and the best possible discount the firm can offer for the contract as θ.

The main primitive is the distribution of θ. To identify it, we make a simplifying assumption that

$$C^P(\theta, q) = 1 - \theta, \tag{10}$$

meaning that there is no connection between past performance q and today's production costs C^P. Since we observe the joint distribution of (s, q), through the optimality conditions (7), we observe the joint distribution of (ρ, q) and, therefore, the marginal distribution of θ is identified.

To identify the joint distribution of $(\theta, \underline{q})$ one also has to assume the shape

[32] See, for example, Che (1993); Asker and Cantillon (2010); Hanazono, Nakabayashi and Tsuruoka (2013); Andreyanov (2019); Laffont et al. (2020)

of investment costs, such as, for example,

$$C^I(\underline{q}, q) = \frac{(q - q_i)^2}{2\beta}, \tag{11}$$

for some $\beta > 0$. However, if the counterfactual is a price-only auction, all revenue-related characteristics will be independent of the exact shape of C^I.

Estimation and simulation

To account for auction-level heterogeneity, as well as a possible evolution of beliefs and strategies over time, we adopt a simple parametrization, where the location of the distribution of discounts (and thus the pseudo-types) is a linear function of the auction-day dummy variables.

Denote the auction-day fixed effects as γ. In other words, if S, D, Q are the observed score, discount, and quality, then

$$S = s + \gamma, \quad D = d + \gamma, \quad Q = q, \quad S = \alpha Q + D. \tag{12}$$

Due to the linear scalability of the optimality conditions, a shift in the location of the distribution of pseudo-types does not affect the shape of the strategy. This motivates (additively) partialling-out the auction-level heterogeneity in a reduced form. We regress the observed score S on auction-day dummies, see Table A.1, to obtain the estimates of fixed effects γ.

Similar to Guerre, Perrigne and Vuong (2000) and Li, Perrigne and Vuong (2000) we use a non-parametric approach to estimate the sample analog of equation (7), see Andreyanov and Franguridi (2021) for details. For any value of β, we can therefore obtain the pseudo-sample of estimated pairs $(\hat{\theta}_i, \hat{\gamma}_i)$, and simulate the outcomes in the counterfactual price-only auction, see Appendix D for details.

Counterfactuals

In this section, we present the counterfactual monetary outcomes for the price-only auction and compare them to the default (with $\alpha = 1/3$) scoring auction, see Table 8. Columns (1) and (3) contain outcomes averaged over all bids. Columns (2) and (4) contain the average winner's outcomes. The

Table 8: Counterfactuals for $\alpha = 0$ and $1/3$

design	scoring ($\alpha = 1/3,\ \omega = 1/4$)		price only ($\alpha = \omega = 0$)	
	all (1)	winner's (2)	all (3)	winner's (4)
quality (%)	87.78 (0.11)	89.92 (0.4)	87.78 (0.11)	87.81 (0.38)
discount (%)	19.48 (0.16)	26.81 (0.3)	19.52 (0.16)	27.03 (0.23)
cost (%)	79.43 (0.16)	71.59 (0.31)	79.43 (0.16)	71.14 (0.26)
markup (%)	1.43 (0.03)	2.26 (0.16)	1.39 (0.03)	2.63 (0.15)
shading (%)	1.09 (0.02)	1.6 (0.11)	1.05 (0.02)	1.83 (0.1)
shading ($)	5470.24 (127.75)	8206.55 (508.14)	5320.39 (126.4)	9349.92 (506.16)
cost ($)	390299.35 (3619.5)	366266.7 (1512.1)	390299.35 (3619.5)	363169.38 (1190.45)

standard deviation is computed via Bootstrap.

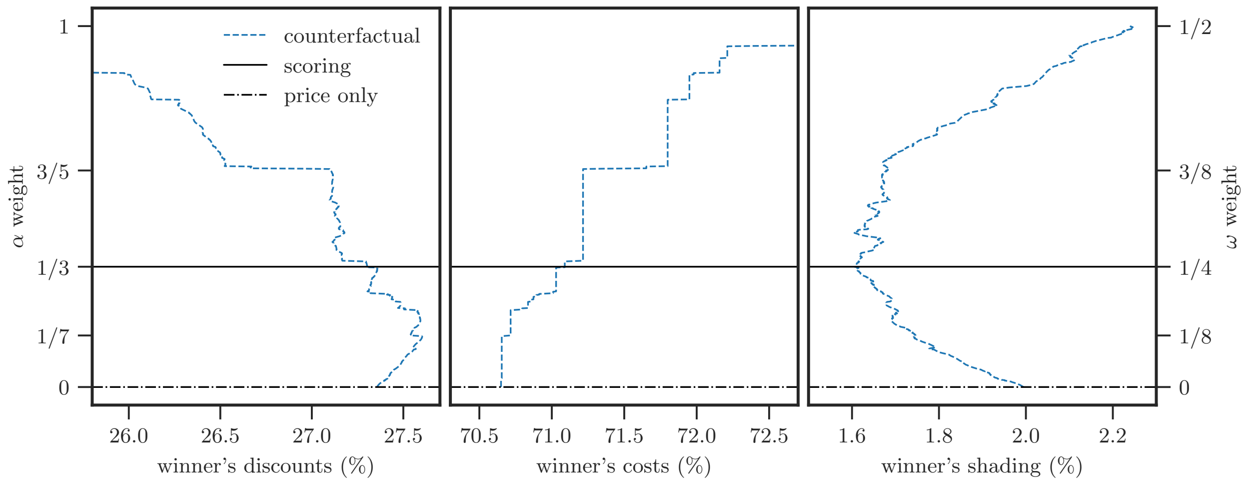
We can see that, relative to the scoring auction, the counterfactual price-only auction is associated with slightly lower (71.13% as opposed to 71.59%) production costs of the winning firm. This is not a surprise since, in terms of production costs, the price-only auction is the most efficient one. However, it is also associated with greater shading (1.83% as opposed to 1.6%), which partially offsets the former. Their combination leads to the price-only auction being slightly cheaper in terms of discounts (27.03% as opposed to 26.81%), but this change is statistically insignificant. We stress that these results do not depend on the choice of β or the exact shape of the cost function.

To put things into perspective, we also simulate counterfactual scoring auctions with other weights, with quality fixed at the level observed in the data. This can be interpreted as a temporary (or unexpected) change in the scoring weight away from $\alpha = 1/3$ or simply a limit when β approaches 0.[33] Thus, for a new scoring weight, α', we only have to calculate the new pseudo-type distribution $F_\rho(\cdot|\alpha')$ and re-evaluate bid shading. We present the results in Figure 8.

The relationship between the winner's expected production costs and the scoring weight is monotonic. This is not a surprise since higher weight means that high-quality firms have an advantage over low-cost firms. However, it is

[33] At the same time, the effect of moving between $\alpha = 1/3$ and $\alpha = 0$ does not depend on β, or the shape of C^I, and can be interpreted as permanent and anticipated.

Figure 8: Counterfactuals for all α



not monotonic for the expected winner's bid shading, similar to what we see in Example 3. This leads to the observed discounts in the scoring auction being very close to those in the counterfactual price-only auction.

Our results indicate that the switch to the scoring auction, through lower shading, has increased quality without a visible increase in the price for the buyer (Acea), which explains the empirical puzzle.

6 Conclusions

This paper has studied the merits of using past performance audits to spur quality and safety in executing public works. The evaluation of the evidence from a reform undertaken by Acea, a large utility company, has shown strong improvements in contract performance after the announcement of its intention to use past performance scores to award future contracts.

Improvements involve all parameters and suppliers, are long-lasting (for at least 10 years after the initial reform), and reflect higher service quality by the utility. Regarding prices, we find mixed empirical evidence and argue that, essentially, all of the improvements in contract execution came at zero or negative cost for the buyer, which is an empirical puzzle.

To explain the puzzle, we employ a novel scoring auction model, which permits such an outcome through the compression of firm markups. The necessary ingredients are (sunk) investment costs and quality concerns (i.e., adverse selection), both inherent to the environment and bivariate types. Our structural analysis also allows us to peek into counterfactual scenarios with scoring weight α other than 0 or $1/3$, see Figure 8.

Several aspects remain open and offer room for future research; for example, how to optimally set the parameter weights, the "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), and how to set the rating for new entrants.

The policy relevance of our findings is significant. There is an ongoing policy debate in Europe and the US on using contractors' past performance 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, see Kelman (1990). Interestingly, the EU follows a very different system, essentially barring the use of past performance except for extremely severe misbehavior sanctioned by the judiciary. Indeed, using 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.[34] To this debate, our results offer a clear empirical illustration of the great potential of a rating mechanism regarding the targeted past performance measures.

[34] Current EU regulation acknowledges the importance of past performance 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 investigator is one of the main selection criteria.

References

Andrews, Isaiah, and Daniel Barron. 2016. "The Allocation of Future Business: Dynamic Relational Contracts with Multiple Agents." *American Economic Review*, 106(9): 2742–59.

Andreyanov, Pasha. 2019. *Essays on Prior-Free Mechanism Design.* University of California, Los Angeles.

Andreyanov, Pasha, and Grigory Franguridi. 2021. "Nonparametric inference on counterfactuals in first-price auctions."

Andreyanov, Pasha, Ilya Krasikov, and Alex Suzdaltsev. 2023. "Scoring and Favoritism in Optimal Procurement Design." Working Paper.

Asker, John, and Estelle Cantillon. 2008. "Properties of scoring auctions." *The RAND Journal of Economics*, 39(1): 69–85.

Asker, John, and Estelle Cantillon. 2010. "Procurement when price and quality matter." *The Rand journal of economics*, 41(1): 1–34.

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.

Bar-Isaac, Heski, and Steven Tadelis. 2008. "Seller Reputation." *Foundations and Trends(R) in Microeconomics*, 4(4): 273–351.

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

Board, Simon. 2011. "Relational Contracts and the Value of Loyalty." *American Economic Review*, 101(7): 3349–67.

Bosio, Erica, Simeon Djankov, Edward L Glaeser, and Andrei Shleifer. 2020. "Public Procurement in Law and Practice." National Bureau of Economic Research Working Paper 27188.

Burguet, Roberto, Juan Jose Ganuza, and Ester Hauk. 2012. "Limited Liability and Mechanism Design in Procurement." *Games and Economic*

Behavior, 76(1): 15–25.

Butler, Jeffrey V., Enrica Carbone, Pierluigi Conzo, and Giancarlo Spagnolo. 2020. "Past Performance and Entry in Procurement: an Experimental Investigation." *Journal of Economic Behavior and Organizations*, 173: 179–195.

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

Camboni, R, L Corazzini, S Galavotti, and P Valbonesi. 2023. "Bidding on Price and Quality: An Experiment on the Complexity of Scoring Rule Auctions." *The Review of Economics and Statistics*, forthcoming.

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, Rodrigo, Andres Gonzalez-Lira, and Michael Walker. 2020. "Rules Versus Discretion in Public Procurement." *mimeo*.

Chassang, Sylvain, and Juan Ortner. 2016. "Collusion in Auctions with Constrained Bids: Theory and Evidence from Public Procurement." *Journal of Political Economy*, 127(5): 2269–2300.

Che, Yeon-Koo. 1993. "Design competition through multidimensional auctions." *The RAND Journal of Economics*, 668–680.

Chong, Eshien, Carine Staropoli, and Anne Yvrande-Billon. 2014. "Auction versus negotiation in public procurement: Looking for empirical evidence." *The manufacturing markets, legal, political and economic dynamics*, 120–142.

Colonnelli, Emanuele, and Mounu Prem. 2021. "Corruption and Firms: Evidence from Randomized Audits in Brazil." *The Review of Economic Studies*, forthcoming.

Coviello, Decio, Andrea Guglielmo, and Giancarlo Spagnolo. 2018. "The effect of discretion on procurement performance." *Management Science*, 64(2): 715–738.

Coviello, Decio, Luigi Moretti, Giancarlo Spagnolo, and Paola Valbonesi. 2018. "Court efficiency and procurement performance." *The Scan-*

dinavian Journal of Economics, 120(3): 826–858.

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.*

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2013. "Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India." *The Quarterly Journal of Economics*, 128(4): 1499–1545.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan. 2018. "The Value of Regulatory Discretion: Estimates From Environmental Inspections in India." *Econometrica*, 86(6): 2123–2160.

D'Alpaos, Chiara, Michele Moretto, Paola Valbonesi, and Sergio Vergalli. 2013. "Time overruns as opportunistic behavior in public procurement." *Journal of Economics*, 110: 25–43.

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.

Guerre, Emmanuel, Isabelle Perrigne, and Quang Vuong. 2000. "Optimal nonparametric estimation of first-price auctions." *Econometrica*, 68(3): 525–574.

Haile, Philip, Han Hong, and Matthew Shum. 2003. "Nonparametric tests for common values at first-price sealed-bid auctions."

Hanazono, Makoto, Jun Nakabayashi, and Masanori Tsuruoka. 2013. "Procurement auctions with general price-quality evaluation." *KIER Discussion Paper*, 845.

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

Holmstrom, Bengt. 1999. "Managerial incentive problems: A dynamic perspective." *The Review of Economic Studies*, 66(1): 169–182.

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

Jones, M Chris. 1992. "Estimating densities, quantiles, quantile densities and density quantiles." *Annals of the Institute of Statistical Mathematics*, 44(4): 721–727.

Kang, Karam, and Robert A Miller. 2021. "Winning by Default: Why is There So Little Competition in Government Procurement?" *The Review of Economic Studies*, forthcoming.

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

Klein, Benjamin, and Keith B Leffler. 1981. "The Role of Market Forces in Assuring Contractual Performance." *Journal of Political Economy*, 89(4): 615–41.

Krasnokutskaya, Elena. 2011. "Identification and estimation of auction models with unobserved heterogeneity." *The Review of Economic Studies*, 78(1): 293–327.

Krasnokutskaya, Elena, and Katja Seim. 2011. "Bid Preference Programs and Participation in Highway Procurement Auctions." *American Economic Review*, 101(6): 2653–2686.

Laffont, Jean-Jacques, Isabelle Perrigne, Michel Simioni, and Quang Vuong. 2020. "Econometrics of scoring auctions." In *Essays in honor of Cheng Hsiao*. Vol. 41, 287–322. Emerald Publishing Limited.

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*, 8(3): 258–83.

Lewis, Gregory, and Patrick Bajari. 2011. "Procurement Contracting with Time Incentives: Theory and Evidence." *The Quarterly Journal of Economics*, 126(3): 1173–1211.

Liebman, Jeffrey B., and Neale Mahoney. 2016. "Do Expiring Budgets Lead to Wasteful Year-End Spending? Evidence from Federal Procurement." *American Economic Review*, 107(11): 3510–49.

List, John. 2006. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions." *Journal of Political Economy*, 114(1): 1–37.

Li, Tong, Isabelle Perrigne, and Quang Vuong. 2000. "Conditionally independent private information in OCS wildcat auctions." *Journal of Econometrics*, 98(1): 129–161.

Lopomo, Giuseppe, Nicola Persico, and Alessandro Villa. 2023. "Optimal Procurement With Quality Concerns." *American Economic Review*, forthcoming.

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

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): 200–249.

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.

Online Appendix

A 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 reform, 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 analysis of the consumer value of the reform, the value of statistical life figures come from the OECD (https://www.oecd.org/environment/mortalityriskvaluationinenvironmenthealthandtransportpolicies.htm), while those for the economic cost of 1 hour of a blackout, separately for business and residential customers come from Table 11 in the ARERA's decision n. 172/07 of 12/07/2007.

B Heinrichs's pyramid

The Heinrichs's (Bird's) pyramid[35] entails the following ratios: 1 fatal accident to 10 serious accidents, to 30 minor accidents, to 600 near misses, to 200,000-300,000 unsafe acts. Assuming that each episode of non-compliance

[35] See Heinrich (1931), Bird and Germain (1986), Goodman (2012) and https://en.wikipedia.org/wiki/Accident_triangle.

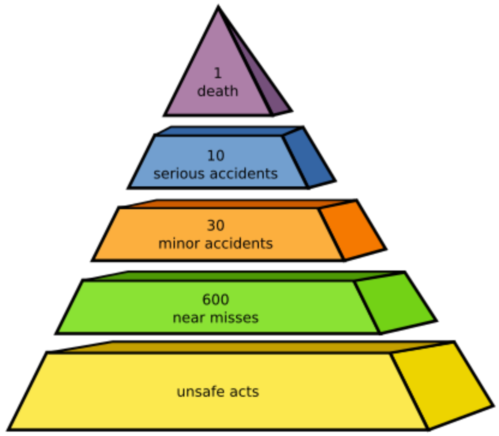


Figure A.0: Heinrich's (Bird's) pyramid.

in the safety parameters audited by Acea corresponds to one unsafe act in the pyramid, we can estimate the reduction in the probability of a fatal accident.

This is calculated as follows: in a typical audit, 33.08 parameters are assessed, 85.3 percent of which are part of the subset of safety-related parameters. There are, on average, 43 contracts a year, with an average duration of 250 working days. If the same rate of compliance observed across audits applies to every working day, then a conservative estimate of a 55 percentage point improvement in parameter compliance implies a reduction in about 163,000 unsafe acts per year. Using the 200,000-300,000 number of unsafe acts, this maps into a reduction in the probability of a fatal accident of 0.54-0.82 per year.

C Theory

C.1 Second order conditions

Note that the firm's action is 2-dimensional. Instead of picking s, q simultaneously, she can optimize over q conditional on s. This is equivalent to plugging equation (8) into the utility:

$$U_i(q(s,\gamma_i),s) = (\theta_i + \alpha(1-\frac{1}{\beta})(\gamma_i + (\alpha G(s))^{\frac{1}{\beta}}) + \frac{\alpha}{\beta}\gamma_i - s)G(s),$$

which she can then maximize over s. The optimal score can be derived via the envelope conditions:

$$s(\theta_i,\gamma_i) = \theta_i + \alpha(1-\frac{1}{\beta})(\gamma_i + (\alpha H(\rho(\theta_i,\gamma_i))^{\frac{1}{\beta}}) + \frac{\alpha}{\beta}\gamma_i - \int_0^{\rho(\theta_i,\gamma_i)} H(z)dz / H(\rho(\theta_i,\gamma_i)).$$

We can then invoke a standard mechanism design argument to show that the second-order conditions are satisfied. Indeed, if the agent reports a score associated with a different type θ' and chooses a quality that is optimal for that score, her utility will be equal to

$$(\theta_i - \theta')H(\rho(\theta',\gamma_i)) + \int_0^{\rho(\theta',\gamma_i)} H(z)dz$$

which has a unique critical point $\theta' = \theta_i$. Finally, the second derivative at the critical point is equal to $-2\frac{\partial H}{\partial \rho}\frac{\partial \rho}{\partial \theta}$, which is strictly negative, thus the second order conditions are satisfied.

C.2 Bivariate types example

For both weights $\alpha = 0, 1$, it is true that $\mathbb{E}\theta|\rho = \rho - \underline{\rho}$. The analytical expression for the expected winning firm's type is, therefore the same:

$$\int \mathbb{E}\theta|\rho dH^2(\rho) = \int 1 - H^2(\rho)d\rho.$$

The expression for the expected winning firm's bid shading, on the other hand, is

$$\int \frac{\int_{\underline{\rho}}^{\rho} H(x)}{H(\rho)} dH^2(\rho) = 2\int H(\rho)(1-H(\rho))d\rho.$$

The expected winning firm's discount is, therefore their difference.

D Structural

D.1 Specification tests

We test whether $s_1 - s_2$ is independent of $s_3 - s_4$, where $\{s_i\}_{i=1}^4$ are four scores, randomly picked in every auction, to validate an additive model of heterogeneity.[36] It is soundly not rejected, according to Pearson ($r = 0.004$ $p = 0.8$) and Spearman ($r = 0.01$ $p = 0.58$) correlation tests with 3000 randomly picked quadruples of scores, see Figure A.1.

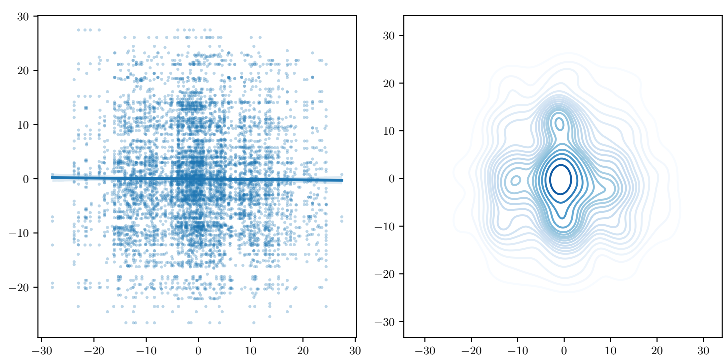


Figure A.1: Scatterplot and contourplot of score differences $s_1 - s_2$ and $s_3 - s_4$.

We apply the analog of the sup-norm test, suggested by Haile, Hong and Shum (2003), to compare the distributions of score residuals between auctions with different numbers of bidders. To test whether the distributions are identical, the statistic is formed

$$\delta = \sum_{10}^{16} \sup_v \{\hat{F}_{n+1}(v) - \hat{F}_n(v)\},$$

where $\hat{F}_n(v)$ is the empirical cdf of score residuals with n bidders. The asymptotic distribution of the statistic is achieved via sub-sampling, and the IPV hypothesis is soundly not rejected ($\delta = 1.44$, $p = 0.52$), see Figure A.2.

[36] Similarly, we could validate a multiplicative model by testing the independence of score ratios s_1/s_2 and s_3/s_4.

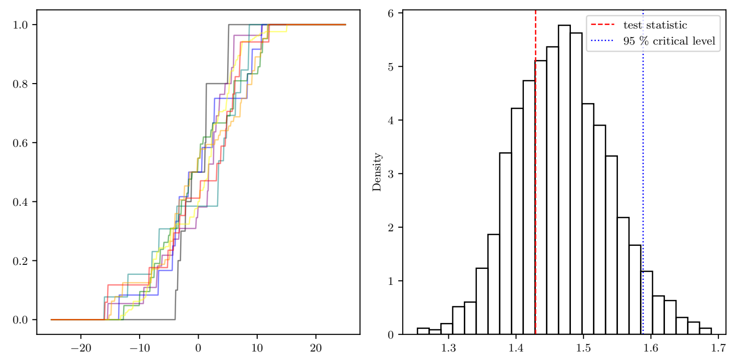


Figure A.2: Empirical CDF's of score residuals (left) with different numbers of bidders and the distribution of the δ statistic (right) obtained via subsampling.

D.2 Quantile approach

We will rewrite the optimality conditions in the quantile form to make the optimality conditions more palatable. For this, we will need auxiliary functions that only depend on the probabilities p_n:

$$C(u) = \sum_{n=1}^{N} p_n u^{n-1}, \quad c(u) = \frac{C(u)}{C'(u)}.$$

Denote $Q_s(\cdot|\alpha), q_s(\cdot|\alpha)$ to be the equilibrium quantile function and density of the score, while $Q_\rho(\cdot|\alpha)$ the quantile function of the pseudo-type. Using the trivial identities $F_s(Q_s(u|\alpha)|\alpha) = u$ and $F_\rho(Q_\rho(u|\alpha)|\alpha) = u$, we can recast the first order conditions as

$$Q_\rho(u|\alpha) = Q_s(u|\alpha) + q_s(u|\alpha)c(u), \tag{13}$$

and the envelope conditions as

$$Q_s(u|\alpha) = Q_\rho(u|\alpha) - \frac{\int_0^u C(x)dQ_\rho(x|\alpha)}{C(u)}. \tag{14}$$

Table A.1: Regression of the score on auction-day fixed effects with (1) additive and (2) multiplicative (target variable is in logarithms) heterogeneity.

	Dependent variable: score	
	(1)	(2)
2011-01-21	3.302 (1.583)	0.327*** (0.034)
2011-01-26	5.734*** (1.915)	0.359*** (0.042)
2011-02-15	-7.209*** (1.831)	0.069* (0.040)
2011-02-16	6.101*** (1.515)	0.369*** (0.033)
2011-02-17	3.068*** (1.152)	0.299*** (0.025)
2011-02-18	5.451*** (1.692)	0.348*** (0.037)
2011-02-22	7.146*** (1.316)	0.389*** (0.029)
2011-02-23	7.103*** (1.160)	0.396*** (0.025)
2011-03-22	2.860 (1.279)	0.294*** (0.028)
2011-03-23	13.682*** (1.583)	0.506*** (0.034)
2011-04-28	-1.973 (0.984)	0.198*** (0.021)
const	45.264*** (0.430)	3.553*** (0.009)
Observations	464	464
Adjusted R^2	0.193	0.184
F Statistic	12.100***	11.410***
Note:		*p<0.1; p<0.05; ***p<0.01

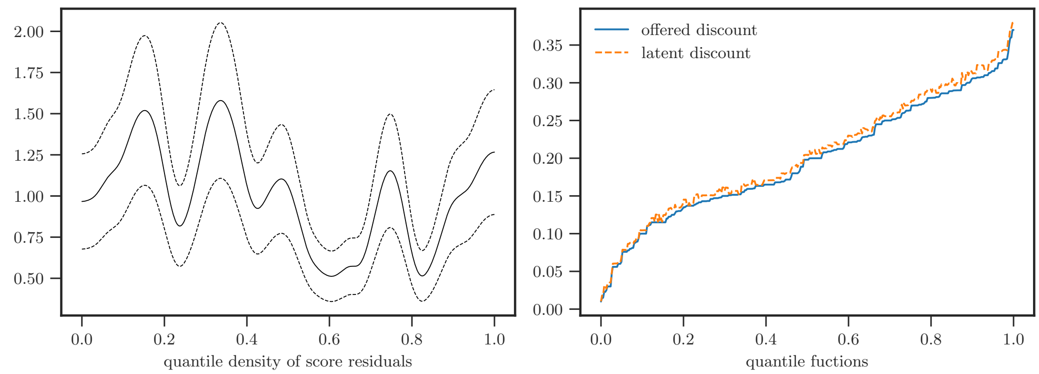
D.3 Estimation

For convenience, we will divide the observed scores by 0.75, so that the default scoring rule has the form $s = \alpha q + d$, with $\alpha = 1/3$.

As is common in the literature, we will first residualize the observed scores to eliminate the auction-fixed effects. Note that this process does not change the ranking of firms within the auction, that is, the firm with the highest residual is also the winner in the data. Note also that while the location of the distribution of δ_i is not identified, it does not matter due to the linear scalability of the model.

The auction fixed effects account for roughly 21% of the variance of the score variable. Denote the residuals and fitted values from the regression as $\hat{s}_m$ and $\hat{\gamma}_m$, $m = 1, \dots M$. We further sort the observations w.r.t. residuals in ascending order and denote the new sample as $(q_{(m)}, \hat{s}_{(m)}, \hat{\gamma}_{(m)})$. We aim at using the identifying equation (13) in order to obtain the pseudo-sample $(q_{(m)}, \hat{\rho}_{(m)}, \hat{\gamma}_{(m)})$, where $\hat{\rho}_{(m)}$ are the estimates of pseudo-types.

Figure A.3: Quantile density and functions



Consider a sample analog of equation (13), evaluated at an evenly spaced grid:

$$\hat{Q}_\rho(u|\alpha) = \hat{Q}_s(u|\alpha) + \hat{q}_s(u|\alpha)\hat{c}(u), \quad u \in \{\frac{m}{M}\}_{m=1}^M. \tag{15}$$

Observe first that $\{\hat{Q}_s(\frac{m}{M}|\alpha)\}_{m=1}^M$ can be thought of as the observed column of (sorted) score residuals $\{\hat{s}_{(m)}\}_{m=1}^M$, while $\{\hat{Q}_\rho(\frac{m}{M}|\alpha)\}_{m=1}^M$ can be thought of as the sought column of pseudo-types $\{\hat{\rho}_{(m)}\}_{m=1}^M$. At the same time, $\{\hat{q}_s(\frac{m}{M}|\alpha)\}_{m=1}^M$ can be obtained as

$$\{\sum_{k=1}^M K_h(\frac{m-k}{M})(\hat{s}_{(m+1)} - \hat{s}_{(m)})\}_{m=1}^M,$$

a non-parametric estimator of the quantile density, suggested in Jones (1992), see Andreyanov and Franguridi (2021) for details. We trim the distribution of residuals at 10% on each end and use a standard combination of a triweight kernel and Silverman rule-of-thumb bandwidth. Finally, $\hat{c}$ can be consistently estimated directly from the data, so the pseudo-sample can be constructed. See the results of estimation in Figure A.3.

While we could, in principle, construct a smooth estimator of $F_\rho(\cdot|\alpha)$ for every α and use it to evaluate each of the counterfactuals, we find it much easier to use the starting pseudo-sample $(q_{(m)}, \hat{\rho}_{(m)}, \hat{\gamma}_{(m)})$ to obtain a counterfactual pseudo-sample $(q_{(m)}, \hat{s}'_{(m)}, \hat{\gamma}_{(m)})$. The counterfactual winner in the auction is, therefore, the firm with the highest counterfactual score $\hat{s}'$.

D.4 Simulations

Consider a sample analog of equation (14), evaluated at an evenly spaced grid:

$$\hat{Q}_s(u|\alpha') = \hat{Q}_\rho(u|\alpha') - \frac{\int_0^{m/M} \hat{C}(x) d\hat{Q}_\rho(x|\alpha')}{\hat{C}(u)}, \quad u \in \{\frac{m}{M}\}_{m=1}^{M}. \tag{16}$$

Again, $\{\hat{Q}_\rho(\frac{m}{M}|\alpha')\}_{m=1}^{M}$ can be thought of as the (nonparametrically estimated) column of pseudo-types, adjusted to reflect the change in the scoring rule:

$$\{\hat{\rho}'_{(m)}\}_{m=1}^{M} = \{\hat{\rho}_{(m)} + (\alpha' - \alpha) q_{(m)}\}_{m=1}^{M}.$$

Furthermore, we can approximate the integral with a sum:

$$\int_0^{m/M} \hat{C}(x) d\hat{Q}_\rho(x|\alpha') \approx \frac{1}{M} \sum_{m=1}^{M} \hat{C}(\frac{m}{M})(\hat{\rho}'_{(m)} - \hat{\rho}'_{(m-1)}),$$

and, of course, the $\hat{C}$ function can be estimated directly from the data. Finally, the counterfactual scores can be obtained as $\{\hat{Q}_s(\frac{m}{M}|\alpha')\}_{m=1}^{M}$ and the counterfactual discounts as $\{\hat{s}'_{(m)} - \alpha' q_{(m)}\}_{m=1}^{M}$.

E Additional Results

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

The estimates in Table A.2 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:

$$Pr(compliant) = \Phi[t + f + \alpha\ weight + \theta\ quick + \gamma_j \sum_{j=2}^{12} category_j], \tag{17}$$

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.

Table A.2: Probability of Compliant Parameter

	<t1				>t1			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weight	-0.024***	-0.022***	-0.022***	-0.023***	0.012***	0.009***	0.006***	0.006***
	(0.003)	(0.005)	(0.005)	(0.005)	(0.000)	(0.000)	(0.000)	(0.000)
Quick		0.075**	0.075**	0.065**		0.015***	0.009***	0.009***
		(0.025)	(0.025)	(0.025)		(0.002)	(0.002)	(0.002)
C2-Documentation		-0.406***	-0.406***	-0.424***		-0.179***	-0.093***	-0.091***
		(0.038)	(0.038)	(0.038)		(0.003)	(0.003)	(0.003)
C4-Works Execution		-0.538***	-0.538***	-0.534***		-0.065***	-0.057***	-0.056***
		(0.042)	(0.042)	(0.043)		(0.003)	(0.002)	(0.002)
C8-Underground works		-0.328***	-0.328***	-0.309***		-0.225***	-0.114***	-0.108***
		(0.034)	(0.034)	(0.035)		(0.003)	(0.002)	(0.002)
C10-Personnel		-0.315***	-0.315***	-0.327***		-0.137***	-0.130***	-0.126***
		(0.048)	(0.048)	(0.048)		(0.003)	(0.002)	(0.002)
C11-Works site regularity		-0.683***	-0.683***	-0.681***		-0.193***	-0.157***	-0.151***
		(0.036)	(0.036)	(0.037)		(0.002)	(0.002)	(0.002)
C12-Works site safety		-0.394***	-0.394***	-0.406***		-0.110***	-0.098***	-0.095***
		(0.037)	(0.037)	(0.038)		(0.003)	(0.002)	(0.002)
Year FEs			✓	✓			✓	✓
Firm FEs				✓				✓
N	3813	3053	3053	3019	361338	256720	256720	256630

Note: 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.

We are particularly interested in the coefficient on *weight* as this has the potential to reveal the strategic nature of supplier responses. Table A.2 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.[37]

Regarding the other coefficients in Table A.2, 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 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.

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

[37] All estimates in Table A.2 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.

Table A.3: Probit Audit Passed

VARIABLES	(1) Evaluation	(2) Evaluation	(3) Evaluation	(4) Evaluation	(5) Evaluation
<t1	-0.899***	-0.616***	-0.618***	-0.750***	-0.659***
	(0.097)	(0.124)	(0.124)	(0.111)	(0.114)
t1	-0.203***	0.017	0.016	-0.154**	-0.077
	(0.034)	(0.081)	(0.081)	(0.061)	(0.065)
t2	0.103***	0.290***	0.290***	0.123**	0.198***
	(0.021)	(0.077)	(0.077)	(0.053)	(0.057)
t3+1	0.669***	0.852***	0.851***	0.666***	0.745***
	(0.024)	(0.077)	(0.077)	(0.057)	(0.061)
t5+1	0.927***	1.093***	1.092***	0.866***	0.941***
	(0.021)	(0.076)	(0.076)	(0.058)	(0.061)
scoring	1.437***	1.547***	1.544***	1.006***	1.070***
	(0.050)	(0.089)	(0.090)	(0.094)	(0.096)
hybrid	1.787***	1.899***	1.897***	1.102***	1.160***
	(0.010)	(0.070)	(0.071)	(0.086)	(0.088)
Win bid - Avg bid	0.000	-0.002	-0.002	0.000	0.004*
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Number of offers			0.001		-0.011***
			(0.002)		(0.003)
Observations	123,173	123,138	123,138	102,516	102,516
Supplier FEs		✓	✓	✓	✓
Auction Controls				✓	✓

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

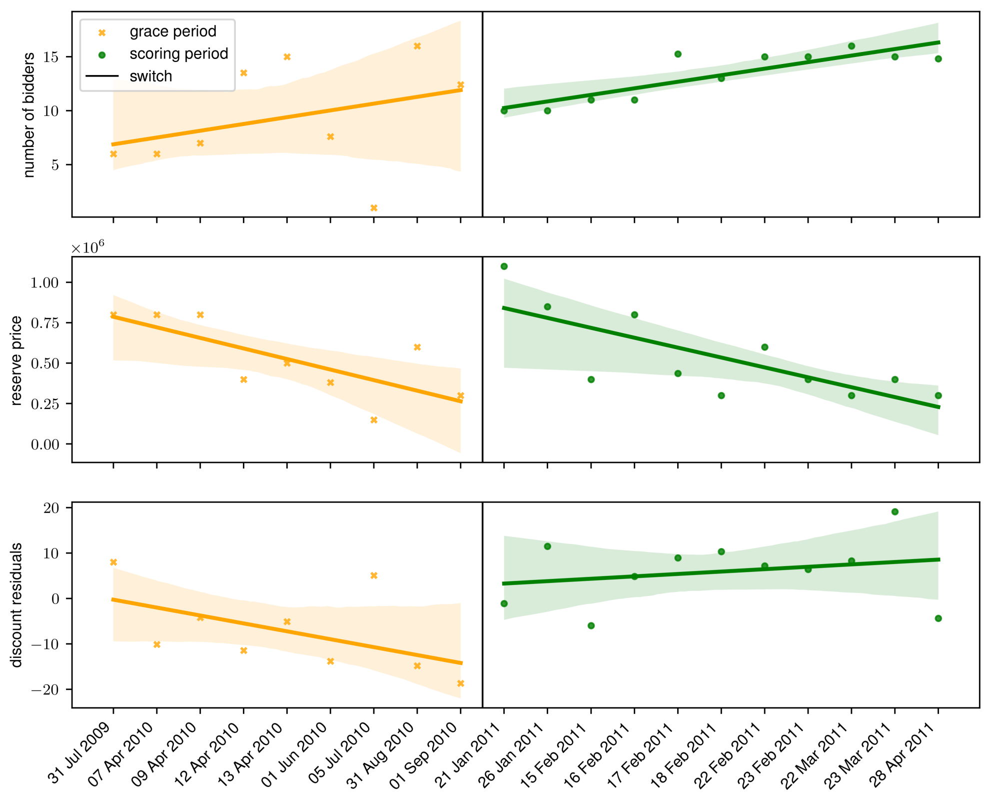
Note: The table reports a series of probit regressions on the probability of passing the audit of a single parameter (1 = passed, 0 = failed). The regressions include several controls, namely: supplier fixed effects, number of offers in the auction, distance between winning bid and mean bid, and contract specific controls.

never expressed to us concerns about episodes of corruption or collusion during the period our data cover.

Table A.3 reports the estimates of a series of probit regressions for the probability of passing the audits in the various periods. Namely: $< t1$ indicates the period right before the $t1$ (i.e., 20 December 2007). $t1$, $t2$, $t3+1$, $t5+1$ are all the breakpoints identified in the time series analysis during the announcement phase (20 December 2007 to 18 May 2010). We progressively control for several confounding factors: the winning bid's aggressiveness (*Win bid - Average bid*), the number of participants in the auctions, a series of contract-specific controls, and firm fixed effects. The time coefficients are generally large and significant, showing an increasing trend from $< t1$ to *hybrid*. The trend is more pronounced when we include firm-specific fixed effects. This shows not only that there is an improvement over time in the scores, but also that the improvement is mostly within the firms.

Finally, Figure A.4 analyzes the jump in discount at the moment of the switch, but controlling for (i.e., partialling out) the number of bidders and the reserve price. The jump of the residualized discount is slightly smaller than in the original exercise, 11 percentage points as opposed to 14, when we average over the 3 auction days before and after the switch.

Figure A.4: Discontinuity at the switch.



Note: Top figure - number of bidders. Middle figure - reserve price. Bottom figure - residualized discounts, with the logarithms of the number of bidders and reserve price partialled-out. Data is averaged by auction day.

F Full list of internal performance measures

Table A.4: Internal Performance Measures

Parameter	Category	Weight
Appliances conditions*	Vehicles	9
Assembly appliances with respect to original design	Cabinet Works	7
Assembly electromechanical equipment	Aerial Works	7
Assembly other equipment	Aerial Works	7
ATM presence*	Documents	10
Bend radius of wires execution	Cabinet Works	7
Binder quality	Underground Works	4
Binder reconstruction - thickness	Underground Works	7
Binding execution	Aerial Works	9
Braces compliant with original design	Aerial Works	5
Braces sealing	Aerial Works	5
Burying material	Underground Works	7
Cabin interferences	Cabinet Works	3
Cleanliness in assembly stages	Joints Exexution	6
Clothing availability*	Works Safety	8
CLS thickness, with respect to prescriptions	Underground Works	7
Columns centering during direct burying	Aerial Works	4
Concession and/or permits*	Documents	4
Concrete transport documents*	Documents	3
Concreting pipe	Underground Works	4
Connection grounding - cabin	Cabinet Works	8
Construction signs*	Works Regularity	4
Correct cable finding	Joints Exexution	6
Correct installation equipotential box	Joints Exexution	7
Correct installation equipotentiality	Joints Exexution	7
Correct schemes continuity recovery	Joints Exexution	7
Display of execution plate	Joints Exexution	5
Disposition DSE(CEL) actuated through notes/minutes*	Works Verifications	9
Document of transport/quality of concrete	Underground Works	8
DPI availability*	Works Safety	10
DPI usage*	Work Execution	10
Drag and deflection	Aerial Works	8
Duct characteristics	Underground Works	4
Duct disposal	Underground Works	4
Electrical connections executions	Cabinet Works	9
Electrical risk checks*	Work Execution	8
Emergency personnel appointment*	Works Safety	10
Emergency personnel presence*	Works Safety	10
Equipotential connection*	Work Execution	10
Extrados height of upper tube	Underground Works	8
Fencing of construction site*	Works Regularity	10
Fencing of deposits*	Works Regularity	5
Fencing of excavations*	Works Regularity	9
Fencing of machine operator*	Works Regularity	8

Fill-in commercial documents	Users Management	6
Filling material compliant	Underground Works	8
Fire extinguisher*	Works Safety	9
Floor plan of the project*	Documents	4
Floor plan of the services*	Documents	7
Following the sequences	Joints Exexution	5
Gas detector*	Works Safety	9
Gas-operated welding instruments	Joints Exexution	5
Graphics*	Documents	5
Groot bed thickness	Underground Works	5
Ground loop compliant with original design	Aerial Works	9
Grounding connection	Aerial Works	9
Grounding of appliances*	Work Execution	10
Grounding of plants*	Work Execution	10
Grounding works compliant with cabinet	Cabinet Works	8
Hydraulic brus-cutter	Joints Exexution	6
Hydraulic press	Joints Exexution	7
Identification*	Personnel	10
Insulated brush-cutting	Joints Exexution	6
Insulating appliances availability*	Works Safety	9
Interferences	Underground Works	7
Interferences	Cabinet Works	7
Interferences - Stretching cables	Aerial Works	6
Material supplies	Underground Works	6
Medical aid kits*	Works Safety	10
Milling - thickness	Underground Works	7
Modification of vehicles and pedestrial circulation*	Works Regularity	9
Observing prescriptions for cable-laying work	Underground Works	7
OTMs conditions*	Vehicles	8
Paintings executions	Cabinet Works	2
Plant delivery documents (PCL)*	Documents	10
Positioning of cross-bars, shelves and so on	Aerial Works	6
Positioning of metal braces	Aerial Works	7
Potential dangers during works*	Work Execution	8
Preliminary notification present and displayed*	Works Verifications	8
Proper clothing usage*	Work Execution	7
Qualifications according to norms CEI*	Personnel	10
Quality of CLS	Underground Works	6
Quality of works	Cabinet Works	4
Realization compliant with original design	Cabinet Works	7
Reels stan*	Work Execution	7
Repaintings executions	Cabinet Works	2
Respect planned meetings	Users Management	8
Sealing ducts in wells	Underground Works	6
Security and coordination plan presence*	Works Verifications	10
Security signs worksite*	Works Verifications	9
Sequences and installation	Joints Exexution	5
Sheet piling	Underground Works	5
Shrinking stages (thermo or auto)	Joints Exexution	6
Sign of machine operator*	Works Regularity	8
Size of excavations	Aerial Works	6
Slope of foundation upper surfaces	Aerial Works	5

Splicing technicians qualified*	Personnel	10
Squareness with axis	Aerial Works	6
Steady polymerization process	Joints Exexution	3
Straight alignment of supports	Aerial Works	6
Subcontractors operating plan presence*	Works Verifications	10
Supplies (cabinet)	Cabinet Works	8
Supplies (I.T.)	Cabinet Works	8
Supply materials 1	Aerial Works	8
Supply materials 2	Aerial Works	8
Supply materials 3	Aerial Works	8
Supply materials 4	Aerial Works	8
Supply materials 5	Aerial Works	8
Support burying	Aerial Works	6
Support positioning	Aerial Works	4
Tent installation	Joints Exexution	6
Timely execution of the works	Users Management	8
Total height	Underground Works	8
Type and quantity of tubes compliant with original design	Underground Works	3
Type of cable	Underground Works	4
Vehicles conditions*	Vehicles	8
Vehicles documents*	Vehicles	10
Vehicles identification*	Vehicles	7
Vertical braces	Aerial Works	6
Visible badge*	Personnel	7
Visual examination of quality and execution	Underground Works	5
Warning signs (night)*	Works Regularity	10
Warning signs (proximity to site)*	Works Regularity	10
Warning signs (vertical and horizontal)*	Works Regularity	9
Warning tape	Underground Works	6
Water tightness verification	Joints Exexution	6
Wear layer reconstruction - thickness	Underground Works	7
Width of excavation	Underground Works	8
Wire stripping 1	Joints Exexution	6
Wire stripping 2	Joints Exexution	5
Workplace cleanliness	Joints Exexution	5
Workplans*	Documents	10
Works awarding*	Documents	9
Works compliant with original design	Aerial Works	7
Works overseers presence*	Works Safety	10
Worksite journal updated*	Works Verifications	7

*Note: Parameters marked with an * are those identified by Acea engineers as most closely related to safety features of the job execution..*