SIEPR Discussion Paper No. 16-036

Contractors' Past Performance and Procurement Outcomes: A Firm-level Experiment

By

Francesco Decarolis, Riccardo Pacini,
Giancarlo Spagnolo

Stanford Institute for Economic Policy Research
Stanford University
Stanford, CA 94305
(650) 725-1874

Contractors' Past Performance and Procurement Outcomes: A Firm-level Experiment

Francesco Decarolis, Riccardo Pacini and Giancarlo Spagnolo*

August 3, 2016

Abstract

Reputational incentives may be a powerful mechanism for improving supplier performance. We analyze their role in contract awarding, exploiting an experiment run by a multi-utility company which introduced a new vendor rating system scoring suppliers' past performance and linking it to the award of future contracts. We study responses in both price and performance to the announcement of the switch from price-only to price-and-rating auctions. Across the 136 parameters scored, overall compliance improves from 25 percent to 80 percent. Improvements involve all parameters and suppliers, but are more pronounced for parameters receiving a higher weight in the new scoring auction. Prices do not significantly change overall. However, we find some evidence of lower prices right after the announcement when firms compete to win contracts to get scored, and of higher prices once all firms have established a good reputation. The experiment suggests that the gains from curtailing suppliers' moral hazard when executing contracts may be higher than those from always bolstering price competition.

JEL: L22, L74, D44, D82, H57.

Keywords: Public procurement, contracts, past performance, reputation, management practices, vendor rating

*Decarolis: Boston University and EIEF, *fdecarolis@icloud.com*. Pacini: Agenzia del Demanio, *riccardo.pacini01@agenziademanio.it*. Spagnolo: SITE-Stockholm School of Economics and CEPR, *spagnologianca@gmail.com*. We are grateful to Nick Bloom, Luis Cabral, Liran Einav, Rosa Ferrer, Hugo Hopenhayn, Neale Mahoney, Enrico Moretti, Gustavo Piga, Piero Sabbatini, Stephane Saussier, Paola Valbonesi and to the participants at the seminars at Barcelona GSE, Berkeley, FGV Brazil, Stanford University, Stockholm University, University of Bologna and University of Paris I for their useful suggestions. We are thankful to the firm running the experiment and to the Italian Authority for Public Contracts for providing us with the data. Decarolis thankfully acknowledges financial support from the ERC (Starting Grant 2015 - 679217).

I Introduction

The importance of reputational forces linking future business to past performance has been documented in decentralized private markets where contracting is limited by the complexity of the traded good or by the lack of legal enforcement.[1] In this paper, we exploit a firm experiment to quantify, for the first time and in considerable detail, the effects of reputational incentives on quality and prices in a centralized public procurement market where law enforcement institutions are available and traded goods are not too complex.[2]

While opinions on the appropriateness of linking past performance to future procurement contracts could not be more divergent across the Atlantic, very little evidence has been available to inform the debate. This is partly because the idea of using reputational scores based on past performance in public procurement is a relatively recent one. In private procurement, past performance indicators have always affected the selection of suppliers and their behavior because private buyers are free to act upon them, e.g. by refraining from selecting suppliers with a poor track record and favoring those with a good one. In public procurement, this type of discretional management practices is typically limited. The need to prevent favoritism and corruption led lawmakers around the world to ensure that open and transparent auctions, where bidders have equal treatment (even when they have a very different track record), are used as often as possible. Another way by which lawmakers limit civil servants' discretion is constraining the use of non-verifiable information (e.g. observable, but non-verifiable qualities of an offer or task performed) to justify supplier selection or monetary transfers. This has often limited the type of reputational scores that public buyers could consider to those based exclusively on verifiable past performance. The reputational mechanism that we study will indeed combine these two characteristics: the use of verifiable measures of past performance within transparent, competitive auctions.

A well known feature of competitive auctions in the context of contract procurement is, however, that, with imperfect contracts, bolstering competition at the bidding stage might

[1] On complex goods, see Banerjee and Duflo (2000) on the Indian software industry; for the lack of legal enforcement, see the burgeoning literature on electronic platforms recently surveyed in Tadelis (2016).

[2] Public procurement refers to the activities through which public authorities purchase works, goods or services from companies.

come at the cost of poor ex post performance. Balancing this price versus performance trade-off, and - specifically - how the use of past performance can contribute to that, is a fundamental, yet unsolved, problem of public procurement. With the Federal Acquisitions Streamlining Act of 1994, the US undertook a major policy change that saw the use of contractors' past performance as the pillar of a new approach to procurement aimed at reducing the rigidity of the procedures built in the Federal Acquisition Regulations. It enabled public buyers to use more flexible and effective purchasing methods similar to private sector management practices, among which is placing a strong weight on suppliers' past performance when selecting bids.[3] Federal agencies thus started to record past contractors' performance evaluations and share them through common platforms for use in future contractor selection.[4] Challenges with collecting and sharing this information created obstacles to the full implementation of this reform, but recent legislation has reinvigorated the role of reputational systems for federal procurement.[5] Interestingly, the EU follows a very different system that exemplifies well the traditional view on past performance in public procurement. Until 2014, the EU Procurement Directives, which coordinate the public procurement regulations of EU member states, essentially forbade the use of past performance, with the exception of extreme cases of major contractual violations sanctioned by the judiciary. This has been one of the features under broad attack during the recent consultation for the revision of the EU Directives, but only minor changes were incorporated in the new 2014 EU Directives: past performance information can now be used to limit bidders' participation in the awarding procedures, but never to select among bidders.[6]

[3]For an early, influential contribution on the comparison of private versus public procurement management practices see Kelman (1990).

[4]The reform was pushed by Steven Kelman (see footnote 3 above) when he served as Administrator of the Office of Federal Procurement Policy in the Office of Management and Budget from 1993 to 1997, playing a lead role in the Administrations "reinventing government" effort that led, among other things, to the Federal Acquisition Streamlining Act of 1994 and the Federal Acquisition Reform Act 1995.

[5]According to a Congressional Research Services report, Manuel (2015): *Reports alleging that contractors received new contracts or orders despite poor performance under prior ones have recently prompted interest in the role that evaluations of past performance play in contracting, as well as attempts by some Members of Congress and the Obama Administration to strengthen requirements pertaining to performance evaluations. (...) the Administration updated the FAR in 2013 to standardize the factors used in evaluating contractors performance, and require that all past performance information be entered into the Contractor Performance Assessment Reporting System. [And to implement the Congress' initiatives] P.L.112-81 and P.L. 112-239.*

[6]See Gordon and Racca (2014) and the responses to the EU consultation at `http://ec.europa.eu/internal_market/consultations/docs/2011/public_procurement/synthesis_document_en.pdf`.

Quantifying the costs and benefits of running a reputational mechanism in public procurement is therefore of first order importance for the sheer economic size of this market,[7] but little reliable evidence is available at present. The main challenge to studying the benefits of using reputation in public procurement is that changes in the way reputation is used are rarely observed. Our study overcomes this problem by exploiting a very rich set of data related to the introduction of a past performance monitoring system in the procurement practices of a large Italian public utility company ("the Firm"). This company is subject to the public procurement regulations, but is also granted some limited flexibility due to its ownership being mixed public-private. In 2007, the Firm started an experiment: to study if it could improve performance through a reputational mechanism, it selected two related types of works in the electricity sector involving the construction or maintenance of either electrical substations or public illumination works. For these works, it laid down a list of 136 observable parameters measuring both quality and safety features of the job carried out.[8] Then, three months after the new audits had begun, the Firm made its first public statement explaining that the results of the new audit system would be converted into a numerical "reputation index" and that, after a few more months of data collection, this index would be used to award new contracts, with a linear scoring rule auction assigning 75 percent of the weight to the price offered and 25 percent to the reputation index.

This paper studies the evolution of both price and performance around the time when the new system was publicly announced, but before scoring rule auctions incorporating reputation were first used. As illustrated in Figure 1, for two and half years after the new audits were introduced, the Firm continued awarding contracts through first price auctions. During this period, the Firm recorded the performance of its contractors and held a series of five meetings with them to explain the functioning of the new system. In the first meeting (indicated as $t1$ in Figure 1), the Firm announced that the new audits would be used in the future to switch from price-only to price-and-rating auctions. In the following four meetings

[7]The OECD estimates that public procurement represents 15-20 percent of GDP across its member countries.

[8]For the Firm, this was considered an experiment in the sense that it introduced the new audit system only for contracts involving two types of works, not for all of the works it procures. The new audit system involved all contracts for electrical substations and public illumination that were already active as of October 2007 or that were tendered after that date.

(indicated as $t2$,...,$t5$), it updated contractors on the forthcoming rule change and it disclosed how compliance with the 136 parameters was evolving across the contractors audited. Our analysis focuses on how price and performance responded to these announcements.

In particular, in the first part of the paper we analyze how compliance in the parameters monitored evolved in response to the timing of the five public announcements the Firm made about the new system. Using audit data for the years 2007-2009, we find clear evidence of a substantial change in contractors' behavior: compliance in the 136 parameters increased from 25 percent before the first announcement ($t1$) to more than 80 percent after the fifth announcement ($t5$). We find that essentially all 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. While this is compatible with a strategic allocation of effort, multi-tasking likely occurred mostly within the set of parameters scored and did not cause a reduction in the effort on unmonitored tasks. In fact, according to the Firm's engineers the broad set of parameters chosen was exhaustive in terms of determining safety and quality for the chosen contract types. Moreover, cost and time overruns - two proxies that are often used for performance but that the Firm decided not to include in the index - did not worsen. Indeed, the Firm's own evaluation of the experiment also found the increase in performance to be fully satisfactory.

The second part of the empirical analysis studies whether the improvement in performance is associated with greater costs for the procurer. This is essential to evaluate the desirability of the switch to the reputational system. This analysis benefits from the nearly ideal timing of the experiment: the continued use of price-only auctions during the period analyzed (see Figure 1) allows us to study price effects linked to performance changes, while ignoring any potentially confounding effect associated with the barriers to entry of an awarding system based on price-and-rating auctions. The empirical strategy used in this part of the analysis takes advantage of a second dataset containing information on all the contracts awarded not only by the Firm, but also by all other Italian public contracting authorities (CAs) during the period 2005-2010. We use the variation across procurers and over time to develop a difference-in-differences estimation strategy. A first set of findings shows that, if

we consider the date of the first announcement to be the one characterizing the occurrence of the policy change, then there is no significant effect on the price paid by the Firm. More specifically, by looking at any symmetric window of time around the first announcement, prices remain stable on average.

However, when we extend the empirical model to account for the evolution of compliance, the price response appears more nuanced. Using the results from the first part of the analysis, we partition the period after the first announcement into a first phase when compliance grows and a second phase when it flattens out at high levels. When we extend the baseline difference-in-differences model to account for these two phases, we find that the original finding of no effect results from the combined effects of prices declining when compliance improves, but increasing after compliance stabilizes. We interpret this evidence as suggestive of a first phase in which suppliers compete harder to win contracts - only contract winners can be audited and, thus, earn or improve their reputation index. That is, winning a contract has the additional benefit of improving the chances of winning future contracts. After all contractors have earned a high reputation index, however, this benefit is outweighed by the increased cost of high compliance, and auction prices become correspondingly higher. The estimates indicate that the overall cost increase of higher quality, after the phase of competition for reputation, ranges between zero and 9 percent of the contract reserve price.

The final part of the analysis studies whether the observed effects are the result of changes in the selection of contractors bidding or in their behavior. The evidence is definitely compatible with the presence of moral hazard: suppliers that are observed bidding both before and after the new rating system is announced stop offering suspiciously low prices. These are precisely the abnormal, low ball bids often associated with poor contractual performance. On the other hand, we find only limited effects of selection based on three features in the data. First, while several suppliers leave the market, the timing of their exit is not associated with the announcements. Second, both the firms that leave the market and those that remain have similar bidding patterns. Third, along many observable characteristics, the firms leaving the Firm's auctions are no different from the firms that leave the auctions of another large multi-utility company that did not participate in the experiment and that we use as a

benchmark. Thus, the main result from this study is that the gains from curtailing suppliers moral hazard when executing contracts may be higher than those from always bolstering price competition, and that a reputational mechanism based on objective past performance can be a powerful tool to achieve this goal.

II Literature

Our paper contributes to a number of literatures on both reputation and public procurement. As mentioned earlier, it contributes to the empirical literature trying to quantify the effects of reputation on quality, prices or contract choice. The already mentioned study of the Indian software industry by Banerjee and Duflo (2000) develops a model of dynamic reputation formation as a signaling game in which each party can propose a type of contract and there is systemic overrun due to the complexity of the goods/services supplied. Empirically, it finds that firms with a better reputation are more likely to be involved in time-and-material contracts and, in most cases, pay for a smaller share of cost overrun. It also finds less firm-generated overrun and less total overrun in contracts involving more reputable firms. This path-breaking study is close to ours in spirit, but at the same time is also rather different. In contrast to that study, we have a fixed contract form and a single public buyer, and we analyze the effects of announcing the introduction of a reputational index based on objectively measured past performance in the scoring rule governing future auctions on detailed measures of quality, safety and price.

The analysis of the effects of the reputational mechanism on prices also connects our study to the literature on reputational mechanisms in electronic platforms recently surveyed by Tadelis (2016). A frequent finding in this literature is that reputation affects the probability of selling, but, as in our study, the effect on price is typically small. Within this literature our paper is probably closer to Klein, Lambertz and Stahl (2016) who, like us, find that an effective reputational mechanism tends to curb in particular moral hazard. Clearly, a major difference is that this literature focuses on reputation based on subjective feedback and on quality outcomes also measured through subjective feedbacks, apart from prices.

Instead, in our experiment we have access to detailed, objective performance measures and a structured reputational mechanism whereby the quality of past performance affects future business directly through the scoring rule (announced) for future procurement auctions.

Our study also contributes to the literature on the impact of competition and dynamic incentives in public procurement with imperfect contracting. Previous theoretical papers have shown that, under imperfect contracting, the results on the optimality of open auctions (e.g., Bulow and Klemperer (1996, 2009)) need not to apply. In particular, Spulber (1990) shows that in the construction sector, where contracting is typically imperfect, open competition spurs moral hazard and ex post opportunism of contractors. More generally, Manelli and Vincent (1995) show that when gains from trade are mainly in non-contractible quality dimensions, open auctions on price only are the worst among all conceivable allocation mechanisms. Bajari and Tadelis (2001) show that bilateral negotiations may be better than open competition for highly complex projects because of the costs of specifying ex ante all contingencies. On the empirical side, only a handful of papers have tried to quantify how procurement design affects both quality and prices. Three recent examples are: Decarolis (2014) on the use of different awarding rules, either lowest price or lowest price with exclusion of abnormally low bids; Liebman and Mahoney (2016) on how yearly fiscal rules on expenditures generate poor quality and prices for goods and services bought close to the end of the fiscal year; and Lewis-Faupel et al. (2016) on how the introduction of electronic auctions lead to improved quality of infrastructure in India and Indonesia.

Our findings are also related to a recent wave of studies highlighting the importance of adopting a dynamic framework to understand public procurement markets. On the theoretical side, that past performance and reputation may play a crucial role for the understanding of repeated public procurement under imperfect contracting was recognized in several studies (e.g. Kim (1998); Doni (2006); Calzolari and Spagnolo (2009), Albano, Cesi and Iozzi (2011)). Overall, this theoretical literature concludes that when contracting is imperfect, (buyer) discretion taking into account past performance can have positive effects on public procurement outcomes. A recent paper by Chassang and Ortner (2016) confirms both theoretically and empirically that a dynamic approach to repeated procurement is indeed

essential to understand the role of "minimum bid requirements" on supplier collusion and procurement outcomes. On the purely empirical side, Marion (2016) studies the interaction between affirmative action programs and firms' capacity constraints in the procurement of US highway construction projects and finds that a dynamic system allowing firms to get exemptions from the requirement to subcontract to "disadvantaged enterprises" by accumulating enough points from previous subcontracts is superior to a static alternative; Gil and Marion (2012) analyze the effect of repeated interaction in the subcontractors market for California's highways and find that past interaction has an effect on bidding behavior only when suppliers expect sizable profits from future interaction; and Coviello, Guglielmo and Spagnolo (2016) show that restricted auctions where only invited suppliers can bid may lead to at least as good outcomes as open auctions, and that when these more discretional procedures are used dynamic incentives are at play: incumbents win more often if they delivered better performance in the past, and they deliver earlier and at lower cost. Our paper contributes to this literature by empirically measuring the power of dynamic incentives induced by a structured reputation mechanism that links current performance to future public procurement contracts by introducing an index of past performance in the scoring rule.

This last aspect also directly links our paper to the literature on scoring auctions in public procurement. In two important recent contributions, Lewis and Bajari (2011) and Lewis and Bajari (2013) use data on US highway construction to show how the high costs that slow highway completion inflicts on commuters can be substantially reduced by curbing moral hazard, optimizing the structure of the procurement contract and inserting time incentives in the scoring rule of the procurement auction. We also show that inserting a measure linked to quality of procurement in the scoring rule can substantially limit moral hazard, but our measure is based on past performance rather than on the performance of the tendered contract. Our evidence that better aligning suppliers' incentives to deliver a high performance needs not to result immediately in higher prices is also found in a recent laboratory experiment on a similar topic by Butler et al. (2014) and in the French public procurement of services by Beuve and Chever (2014). The latter study is close to ours as it empirically studies price and performance responses in 102 contracts for cleaning services where a public buyer, in response to a court ruling invalidating its use of past performance

to exclude a bidder, decided to invest in greater contract completeness.

The particularly detailed performance measures from random audits by centrally managed inspectors we have access to relate our paper to Olken (2007)'s study of corruption in Indonesia, where similarly detailed performance measures are used to evaluate the effectiveness of this kind of random audits against corruption. Related to the issue of corruption, Bandiera, Prat and Valletti (2009) show that, in the context of the experiment that we study, corruption concerns could be less of a priority than finding ways to increase performance, e.g. with reputational mechanisms. Their paper studies waste in the procurement of Italian goods and manages to quantify active waste linked to deliberate corruption and passive waste linked to inefficiency, incompetence, red tape, etc.. It offers important, empirically grounded indications to policy makers. Our paper is a step in the same direction, as it identifies and quantifies the likely costs, benefits, and channels of a change linked to the introduction of a reputational mechanism based on objective performance measures in public procurement.

Finally, on the policy side, this study makes an important contribution to the above mentioned debate. The slow adoption in US federal procurement of the 1994 policy reform on reputational mechanisms suggests that its implementation costs are not trivial. Our paper provides the first empirical evidence on the likely benefits of implementing a reputational mechanism in public procurement, suggesting they may be rather large. This is important in the European context too because the reliance on past performance could reduce the worry, expressed by several experts (see, for instance, Saussier and Tirole (2015)), that the ongoing shift toward a discretionary system of awarding and renegotiation procedures - under the new EU Procurement Directives of 2014 - will create distortions in the EU public procurement market.

III The Context of the Experiment

The experiment entailed the introduction of a new vendor rating system by a large procurer, "the Firm". This is one of the largest public multi-utility companies listed on the Italian stock exchange. The Firm operates in the sale and distribution of energy, water services and

public lighting.[9] In order to maintain an orderly functioning of its power grid, each year the Firm outsources works worth over €300 million. Since the Firm is controlled by a majority shareholder that is a local administration, the procurement of these works must follow the awarding rules laid down in the Italian Public Procurement Code ("the Code").[10]

Being a multi-utility company, the Firm falls witin the "special sectors" which enjoy some flexibility in applying the Code. Starting in 2007, this allowed it to begin an experiment with a new vendor rating system. The basic idea was to set an objective measure of contractual performance, with the plan of using its ratings in the awarding stage of future procurement processes. Below, we first describe how the reputation measure was constructed and then how it was incorporated into the auctions. The latter step faced a series of legal obstacles that we describe at the end of this section and that crucially affected the timing of the experiment. In particular, they induced the Firm to slow down the implementation of the announced switch to price-and-reputation auctions. The timing of the experiment, as well as the time span of our analysis sample, is shown in Figure 1.

A. The Reputation Index

The Firm designed an experiment with a vendor rating system focusing on the procurement of works in the electricity sector. It selected two types of contracts, involving either public illumination or electricity distribution (mostly entailing the maintenance of electrical substations and wires), that were considered sufficiently homogenous to define a list of items key to assessing contractual performance. A total of 136 parameters were identified for this goal. As shown in Figure 1, beginning in October 2007 all new and ongoing contracts for public illumination or electricity distribution began to be (randomly) audited for these parameters. As Table 1 reports, the set of 136 parameters is divided into 12 categories, further divided into 2 macro classes: "safety" (51 parameters; 7 categories) and "quality" (83 parameters; 5 categories). For instance, "Equipment and machinery," the first category in Table 1, comprises 5 parameters involving the adequacy of both the formal documentation

[9] In 2010, the Firm had a turnover of €3.6 billion and produced 15.651 GWh of electricity, making it the sixth largest operator in Italy.

[10] The Code, Legislative Decree 163 of 12 April 2006, is the law that implemented in Italy the European Union public procurement directives 17/2004 and 18/2004 and that was relevant for the period that we study.

and the physical conditions of equipment and machineries.[11] Parameters in this category are quite general and can be inspected for all work sites. Other categories, however, involve parameters specific to a subset of contracts only. For instance, the 25 parameters in "Underground works" involve features that are assessed only for jobs involving underground wires and electrical substations.

[INSERT TABLE 1 APPROXIMATELY HERE]

The system works as follows. Scores are collected by teams of rotating auditors (Firm engineers) in one or more visits to the work sites, with a score assigned to each of the 136 parameters. The score is 1 if the value is "compliant," zero if "not compliant" or "n/a" if it is impossible to inspect. Which contracts are audited and which engineers from the Firm are assigned to the team inspecting each work site are both determined through a process of random drawing. Thus, a single contract might be audited one or more times and by the same or different engineers.

The scores on the individual parameters are then aggregated in a unique reputation index (RI). Each parameter is associated with a weight, ranging from 2 to 10, and the RI is calculated as a weighted average mean across a predefined time span according to:

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

with $p_{ij} \in \{0, 1\}$ indicating the score obtained in each of the n parameters over all the m audits considered and with $u_j \in \{2, 3, ..., 10\}$ being the weight attached. Hence, the reputation index can range from 0 to 1 and entails no differential discounting of audits taking place within the predefined time window. In the period that we study, this window was announced to be equal to one year.[12]

[11] While clearly important for the safety of the work site, these features also influence the quality of the work executed, thus making the distinction between the two aggregate classes of quality and safety rather blurry. Indeed, this is also the opinion expressed to us by representative at the Firm. Therefore, we will make only minimal use of this distinction in our analysis.

[12] More precisely, if the tender is announced in month $n = \{1, ..., 12\}$, then the RI used all audits collected in the previous 12 months starting from $n - 2$, if the announcement take place before the 10th day of the

Finally, it is worth mentioning that past performance was recorded even before the RI system was introduced. Before then, engineers used to inspect work sites and write descriptive memos about their conditions. These memos, however, did not translate into any quantitative assessment of performance and they only served to keep track of how the work was evolving.[13]

B. The Scoring Rule Auction

The regulations in the Code require the Firm to award its contracts through auctions. Incorporating the RI in the awarding process required a switch of the auction award criteria from the lowest price to the most economically advantageous tender (MEAT). As shown in Figure 1, on December 2007, three months after the introduction of the new auditing system, the Firm announced to its contractors its intention to switch to MEAT by adopting a linear scoring rule auction whereby the contract is awarded to the firm with the highest score S calculated as:

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

where w_{price} is the weight assigned to price relative to that assigned to the RI. The firm held five meetings, marked as $t1$,...,$t5$ in Figure 1, to demonstrate the new system to its contractors:

- At the December 2007 meeting ($t1$), the contractors were given a detailed presentation of the forthcoming reform and it was explained how, from a status quo of a $w_{price} = 1$, the new system would have had $w_{price} = 0.75$. Numerical simulations were presented to demonstrate the benefits of accumulating a high RI. Furthermore, the Firm announced that the RI, calculated as above, would apply exclusively to those bidders audited at least 7 times in the relevant time window.[14] Otherwise, a bidder would be assigned an RI equal to the average RI of the bidders in the auction. The same averaging rule applies for new entrants (i.e., firms never audited).

month, or $n-1$, if it it takes place after it. If the announcement takes place between July 10th and September 10th, then the 12 months are considered up until July 10th.

[13] Although they could have been used to enforce penalties, penalties are rarely enforced in this market.

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

- During each of the following four meetings held between April 2008 and January 2009 ($t2$,...,$t5$), the Firm gave updates on the functioning of the vendor rating system, confirming what was announced in $t1$ and showing how the recorded compliance was evolving; the evolution of compliance on individual parameters across each contractor audited was disclosed. The latter occurred without disclosing the contractors' identity, as the Firm agreed with its suppliers to keep their past performance data confidential. Further numerical simulations of the use of the RI in hypothetical scoring rule auctions were also presented.

Thus, contractors' incentives had likely changed already at $t1$, well before the first scoring rule auction was implemented in May 2010. At $t1$, suppliers learned that all contracts had become dynamically linked through the RI formula. Moreover, at each of the $t2$,...,$t5$ announcements they also learned their competitive situation relative to the other suppliers in terms of the RI accumulated. This is suggestive of potential changes in strategic entry and bidding choices that our analysis will seek to uncover.

Finally, it is worth emphasizing that suppliers should have not expected any change in the reserve price relative to the pre-$t1$ phase. This is because the Firm is not in full control of the reserve price. This quantity, publicly known to suppliers at the time of bidding, is obtained by multiplying input quantities (estimated by the Firm's engineers) by their prices and summing up these products. Crucially, input prices are not the current market prices but the list prices set every year by the region where the Firm operates and used exclusively by contracting authorities to calculate reserve prices.

C. Legal Limits and the Timing of the Experiment

Several features of the experiment described above, including the types of measures entering the RI, the use of a scoring rule auction and the slow transition to the latter, are all closely linked to the legal institutions within which the Firm operates. Without entering too much into the intricacies of the legal system, it is worth mentioning that the use of reputational elements for the selection of contractors in public procurement has received widespread attention in the drafting of the EU Procurement Directives. Directive 18/2004 required that

"contracting authorities shall treat economic operators equally and non-discriminatorily and shall act in a transparent way"[15] and that competition is a pillar of the procurement system.[16] For special sectors, however, Directive 17/2004 was less stringent, since it allowed public buyers to institute their own qualification system or, in general, to select potential candidates to be awarded on the basis of their technical and professional skills, chosen at the discretion of the contracting authorities. The unique limit in the choice of such criteria is objectivity: "contracting entities which select candidates for restricted or negotiated procedures shall do so according to objective rules and criteria which they have established and which are available to interested economic operators."[17] Thus, reputation indicators can be used if based on measurable parameters that are verifiable by third parties and agreed upon by contractors. These features shaped the choice of the parameters in the RI.

The problem arises in the awarding phase. Since the EU gives special prominence to the free and fair competition principle, the use of reputation as an award criteria in public procurement can constitute an unfair advantage for the incumbents and a disproportionate disadvantage for new entrants: a potential supplier with no past experience cannot enjoy any reputational premium with respect to pre-existing competitors. This may reduce entry and competition and violate the general principle of equal treatment.[18] In the contract awarding phase, the MEAT[19] is the criterion that allows criteria other than the price to be considered. The EU Court of Justice, however, ruled that the awarding authorities, when evaluating quality with the MEAT, should consider the object of the tender and not the bidder's characteristics.[20] The Italian Procurement Authority reaffirmed the same principle.[21]

[15] Art. 2 of Directive 2004/18/EC of 31 March 2004 on the coordination of procedures for the award of public works contracts, public supply contracts and public service contracts.

[16] "Contracts should be awarded on the basis of objective criteria which ensure compliance with the principles of transparency, non-discrimination and equal treatment and which guarantee that tenders are assessed in conditions of effective competition" (Recital 46, Directive 2004/18/EC). "Non-discriminatory criteria should be indicated which the contracting authorities may use when selecting Competitors and the means which economic operators may use to prove they have satisfied those criteria" (Recital 39, Directive 2004/18/EC).

[17] Art. 54 comma 2, Dir.17/2004/CE.

[18] Clearly, how to treat new entrants is purely a designer choice (see Butler et al. (2014) for a discussion). However, the description of all reputation-based systems as a form of incumbency advantage is how the policy debate has framed the problem.

[19] For awarding criteria specification, see art. 53 Dir.2004/18/EC and art. 55 Dir. 2004/17/EC.

[20] See judgments in Causes C-488/01 or C-31/87.

[21] See AVCP Resolution n. 30 of 06/02/2007.

These restrictions on the use of the MEAT had major impacts on the timing of the experiment. First of all, while the Firm was determined to use the RI to improve contract performance,[22] the risk of being accused of violating the EU Court of Justice ruling on the MEAT was perceived as a serious threat. This initially caused delays in the switch to the scoring rule of about two and half years. Then, after only a dozen scoring rule auctions had taken place between the second half of 2010 and the beginning of 2011, a new management team took charge of the Firm. It opted for a more conservative interpretation of the rules and returned to price-only auctions, maintaing the system we described above for monitoring purposes only.[23]

IV Data

The analysis is based on two sets of data: audits data, covering the performance recorded through the new auditing system; and auction data, covering bidding and other auction-related information. The data on audits, in particular, represent a rather unique opportunity to observe contractual performance. Their major limitation is that a link with the auctions dataset cannot be established because of anonymity requirements imposed by the Firm.[24]

A. Audits

The "Audits data" is a panel dataset that contains the outcomes of all the audits performed between the introduction of the new auditing system, on October 16, 2007, and November 19, 2009. The Firm provided us with this dataset, which contains 64,537 observations recording the scores assigned to each of the 136 parameters inspected during 1,951 audits involving 187 contracts and 45 different contractors. Table 2 reports some summary statistics, aggregating

[22]Some data and experiences show that penalties are not effective because they are not even applied: a study conducted for Consip, the Italian public procurement agency, on a sample procurement contracts on goods and services, demonstrated that penalties were applied in just 3.7% of eligible cases.

[23]Its usage to set a minimum threshold to admit bidders to the auction is currently under consideration by the Firm. As discussed earlier, this is the only usage of past performance that is explicitly allowed under the 2014 EU Procurement Directives.

[24]While the awarding of a public contract is considered information that must be accessible to the public to limit corruption risks, the performance of contractors is considered sensitive information. An analogous distinction is present in the US regulation as well, so that, in accordance with FAR 42.1503(4)(d), information that is stored in the Past Performance Information Retrieval System (PPIRS) is classified as Source Selection Sensitive and is not releasable unless directed by the agency who submitted the data.

parameters at the level of the 12 categories. The table shows that there is substantial heterogeneity in how many times the parameters in each category are inspected. It also reveals that about 20 percent of the observations involve public illumination (PI) works, with the rest being electricity distribution works. The last three columns of the table report the share of compliant parameters, dividing the sample period in three phases: pre $t1$, post $t5$ and between these two phases. For nearly all categories there is an improvement over time: for instance, for "Work site safety" the average pre $t1$ compliance is 30 percent, between $t1$ and $t5$ it is 61 percent and post $t5$ it is 81 percent. While the size of the increase differs across categories, Figure 2 confirms that overall, the increase in the average compliance grows from about 25 percent to 80 percent. The average compliance measure in this figure is calculated by averaging together the 0/1 scores assigned in all audits taking place during the month of reference and weighting them by the weights used in the RI formula. The vertical bars show the 95 percent confidence interval for the mean. Their size tends to decrease over time, due to less variance in the recorded compliance. The next section explores the timing and sources of the evidence shown in Figure 2 .

[INSERT TABLE 2 APPROXIMATELY HERE]

[INSERT FIGURE 2 APPROXIMATELY HERE]

B. Auctions

The second dataset contains data on the awarding of public procurement auctions held between 2005 and 2010 for the type of works involved in the Firm's experimentation. The data covers the Firm as well as all other Italian public contracting authorities and its source is a private company, Telemat spa, which provides information on both past and perspective public tender to firms subscribing to its services. The data include the object of the contract, the reserve price, the awarding price and date, the identity of both the procurer and the winning contractor, and various other contract-specific information. For a subset of auctions, we integrate the data with the information on the losing bids and on the subsequent life of the contracts.[25] Table 3 reports summary statistics dividing the auction dataset into four

[25]For this additional information we have three sources. The Firm gave us access to the full set of bids -

subsets: auctions held before or after $t1$; and held either by the Firm or by other procurers. The comparison of the top and bottom panels on the left side of Table 3 shows that the average winning discount in the Firm's auctions slightly declines, from 22.8 percent to 20.6 percent, but not in a statistically significant way. A similar conclusion is suggested by the visual inspection of the scatter plot in Figure 3 which shows a lack of any clear pattern across the winning discounts in the auctions held by the Firm. However, for the post-$t1$ period, the final phase of the sample contains slightly lower discounts than those present right after $t1$. The following analysis will study this aspect in depth.

[INSERT TABLE 3 APPROXIMATELY HERE]

[INSERT Figure 3 APPROXIMATELY HERE]

The difference-in-differences analysis that we will present in the next section will broadly confirm this overall lack of clear price changes associated with $t1$. For this analysis, we will exploit the presence of control group auctions. These are auctions similar to those of the Firm, but held by different contracting authorities. For the DD analysis, it is important that the auctions in the control group are sufficiently comparable to those held by the Firm. In this regard, Table 3 shows that the winning discount, the contract duration and the share of public illumination contracts is quite similar across the two groups, but that the awarding price tends to be higher in the Firm's auctions. Nevertheless, it is important to stress that the main effort to ensure the comparability of the auctions was at the data collection stage, where we selected only auctions that, in terms of their object, were a close fit for the public illumination and electricity distribution contracts auctioned off by the Firm.[26] Given the comparability of the winning discounts across the two groups, it is interesting to graphically compare the evolution over time of this variable for both groups (see Figure 4). Before $t1$ the

both winning and losing bids. For other contracting authorities, we obtained the same information through textual analysis of the official documents of the contract award. Finally, information on the subsequent life of the contracts, including time and cost renegotiations, was available from 2005 up to the first half of 2008 through the dataset of the Italian Authority for Public Contracts, which covers the universe of all public procurement auctions with a reserve price of at least €150,000.

[26]These works belong to a well defined contract category identified by the Italian regulation as $OG10$, which makes it feasible to select comparable projects.

two series have very similar trends, supporting the idea of using a DD analysis. Moreover, Figure 4 offers a first illustration of the different behavior of winning discounts that the Firm faced. Discounts increased right after $t1$, but then, roughly after $t5$, they substantially decreased. The following analysis tries to establish these effects more formally and to offer an interpretation for them.

[INSERT Figure 4 APPROXIMATELY HERE]

V Empirical Analysis

This section analyzes the effects on price and performance of the Firm's announcements. We begin from an assessment of performance using the audits data. We then move to an analysis of prices exploiting the auctions data.

A. Effect of Announcements on Performance

Figure 2 shows a marked improvement in compliance during the sample period. The three questions that we explore here are: can this increase be associated with the timing of the announcements? Is its magnitude confounded by composition effects? What does it reveal about suppliers' behavior? Regarding the first question, Table 4 reports the results of Chow and Bai-Perron tests for the presence of structural breaks in the time series of the compliance measure. As for Figure 2, the variable analyzed in the first two columns of the table is the monthly weighted average compliance across all parameters. The next two columns restrict the parameters to those in the quality class, while the last two columns use the subset of parameters in the safety class. The top panel of the table reports the results of a Chow test for the presence of one break at $t1$ (odd numbered columns) and five breaks, at $t1, ..., t5$ (even numbered columns). For all the six tests, we reject the null of no breaks in favor of the alternative of breaks at the specified dates. More interesting, however, are the outcomes of the Bai-Perron tests, where we do not specify the dates of the breaks and instead let the test determine them, either without specifying how many breaks there are (odd numbered columns) or specifying that there are 5 breaks at unknown dates (even

numbered columns). The test results are a clear indication that $t1$ is a breakpoint. This confirms that the first meeting held by the Firm with the suppliers to explain the RI system had a significant impact on their contractual performance. As regards the other break dates, all tests allowing for an unspecified number of breaks identify a break near $t5+1$.[27] This is also quite revealing since, by the fifth meeting, suppliers found out that average compliance had reached a fairly high level across all suppliers and parameters. As discussed below, this likely changed the strategic environment in the auctions, through a change in the perceived value from further improvements in compliance.

[INSERT TABLE 4 APPROXIMATELY HERE]

The second question that we explore is whether the higher compliance observed is the result of different sets of parameters or contractors audited. We begin by looking separately at parameters in the quality and safety classes. Figure 5 reports for each month the total weight (averaged across all audits in the month) of parameters relating to these two classes. Safety parameters always carry a higher total weight, but their proportion relative to the quality parameters remains rather stable over time. Indeed, the evolution of the monthly average weighted parameters in these two classes reported in Figure 6 confirms a clear upward trend for both of them. As the latter four columns of Table 4 show, breaks in both series occur at $t1$, but the dates of the other breaks are not all identical. This is also related to the speed of adjustments in compliance, as we will discuss below. Before that, we complete the graphical analysis of the composition issue by taking an even more disaggregated view of the performance measure through their grouping into categories. This is particularly relevant because, as the examples in section 2 showed, the distinction between the safety and quality classes is blurry. As Figure 7 shows, the number of parameters audited per month is quite heterogenous, but Figure 8 reassures us that the increase over time is quite homogenous across categories.[28] Similarly, while there is heterogeneity in how many audits each contractor receives,[29] performance increases are rather homogenous across contractors.

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

[28] To make the figure easier to interpret, we reported only the 4 most audited categories, but the increase is present essentially in all 12 categories, as also revealed by the summary statistics in Table 2.

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

Figure 9 reports the monthly weighted average compliance by separating contractors into four groups on the basis of the quartile of the distribution of the number of audits they receive.

[INSERT FIGURES 5, 6, 7, 8 and 9 HERE]

The final question concerning performance is to what extent suppliers responded to the nuances of the new system. In particular, as sometimes observed in experiments, the mere change in the environment might trigger some response, or, more specifically, as a form of *Hawthorne effect* (or observer effect), suppliers might improve performance once they are aware of the new monitoring.[30] Alternatively, behavior might not be improving at all and what we observe may reflect contractors beginning to collude with their monitors. To address these concerns, we resort to a series of probit regressions performed at the level of each individual audited parameter. In particular, we run the following probit regression for the probability of the score being 1 (i.e., compliant) on features of parameters, contracts and suppliers:

$$Pr(compliant) = \Phi[t + f + \alpha\ weight + \theta\ quick + \gamma_j \sum_{j=2}^{12} category_j], \tag{3}$$

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.

[INSERT TABLE 5 APPROXIMATELY HERE]

We are particularly interested in the coefficient on *weight* as this has the potential to reveal the strategic nature of the suppliers' responses. Table 5 shows the probit marginal effects for two separate samples: audits held in the period before $t1$ (first four columns), and

[30] An *Hawthorne effect* is a change, typically an improvement, in some aspects of behavior in response to the awareness of being observed.

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 on 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.[31]

Regarding the other coefficients in Table 5 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 with 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 the "personal protection tools" (mostly helmets) or the presence of signs warning of the presence of ongoing works nearby. The adequacy of the machineries, instead, is an example of a parameter with *quick* equal to zero. While clearly arbitrary, this dummy variable is helpful to test the reasonableness of the performance response observed in our data. 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 regards to the behavior in the auctions. However, it is relevant here that while it is impossible to fully rule out the possibility of collusion/corruption, the system of random rotation of auditors and of random selection of the sites to inspect was explicitly meant to curtail these types of risks.

B. Announcements Effect on Price

The effect of the announcement on prices is analyzed through a difference-in-differences (DD) strategy. The unit of analysis are the auctions held by the Firm (treated group) and

[31] All estimates in Table 5 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.

by other CAs (control group). To identify the causal effect of the Firm's announcement at $t1$ on prices, we estimate the following regression model:

$$D^w_{ist} = a_s + b_t + cX_{ist} + \beta_1(Treatment) + \epsilon_{ist}, \tag{4}$$

where D^w is the winning discount (over the reserve price) and the index i indicates the auction, s the entity awarding the contract and t the year. $Treatment$ is a dummy variable equal to one for the contracts awarded by the Firm from $t1$ onward and zero otherwise. The coefficient of interest is β_1, the effect of the announcement on the winning discount, conditional on fixed effects for the entity awarding the contract (a_s) and time (b_t), and on other covariates (X). The latter set includes characteristics of both the contract - four dummy variables for value of the reserve price and two dummy variables for the work type - and of the contracting authority - its region and whether it is a local authority (municipality, county or region) or not.

In addition to the break at $t1$, we also exploit the second break detected by the Bai-Perron test at $t5+1$. This allows us to account for the two differential phases of accumulation of RI and stabilization of RI. Thus, we extend the previous model to include a dummy for auctions held from $t5+1$ onward, $D_{t>t5+1}$:

$$D^w_{ist} = a_s + b_t + \beta_1 Treatment_{st} + \beta_2 Treatment_{st} * D_{t>t5+1} + D_{t>t5+1} + \gamma X_{ist} + \varepsilon_{ist}, \tag{5}$$

Under this second model, β_1 now measures the effect on the Firm's awarding discounts past $t1$, but before $t5+1$, while β_2 measures the same effect for being after $t5+1$, relative to the $t1$ to $t5+1$ period. Hence, the effect of the RI accumulation phase is captured by β_1, while that of the RI stabilization phase is captured by β_2.

The identification of the key parameters in the two models above crucially hinges on the validity of the auctions in the control group to capture price variations that would have affected the Firm's auctions absent its reform. Specifically, while Figure 3 suggests that discounts did not change around $t1$, this might be due to simultaneous changes in market conditions. Our Auctions data covers similar contracts awarded by all Italian public

procurers. Thus, the discounts in the control group auctions have the potential to capture price variations at the market level. Figure 4 is indeed reassuring of the fact that the similar pre-$t1$ dynamics in the treatment and control auctions make the parallel trends assumption likely to hold. Therefore, we proceed by first presenting our baseline DD estimates and then exploring their robustness to both identification and inference concerns.

Table 6 presents these baseline estimates for the models in equation (4) (first three columns) and in equation (5) (last three columns). For each model, estimates for three specifications differing on the set of covariates, X, are presented: we first include in X only the constant, then add procurer characteristics, and - finally - also add contract characteristics. We present results for two control groups: estimates in panel (a) use as all contracting authorities, while those in panel (b) use only procurers in central regions. By restricting the attention to contracts held in the same geographical area in which the Firm operates, the latter control group has greater potential to capture price variations due to local conditions. However, it also has greater potential for contamination, as we explore below.

[INSERT TABLE 6 APPROXIMATELY HERE]

The results in the top panel of Table 6 show the lack of any price effect when the post-$t1$ period is considered altogether (first three columns). In addition to not being statistically significant, the estimated coefficients are relatively small in magnitude, implying a 4 percent decline in discounts, when compared to the major shift in performance documented above. Interestingly, the estimates change, revealing a rich price dynamic if the post-$t1$ period is divided into a phase pre and post $t5+1$. The estimates in the last three columns confirm the visual evidence of Figure 4: discounts initially increase, by about 6 percent of the reserve price, and subsequently decline, by about 15 percent of the reserve price. All estimates are highly statistically significant. This also implies that the discounts after $t5+1$ are 9 percent lower than pre $t1$. This effect results from the difference between β_2 and β_1. The bottom panel of Table 6 shows that, despite the smaller sample, very similar results, in terms of size and significance, are obtained with the second control group of auctions held in central Italy.

The remaining part of this section explores the robustness of Table 6 estimates to three

types of concerns. First, the experimentation began by the Firm might feed back to auctions held by other contracting authorities. This contamination of the control group auctions might occur through changes in either contractors costs or market structure. For instance, since some of the parameters scored involve durable equipment and machineries, the investments made by the Firm's contractors might also alter their costs in the auctions held by other CAs. Market structure might also change if some of the Firm's contractors respond by participating more or less in auctions held by other CAs. Since the contamination effect, if present, is likely driven by the presence of common contractors between the Firm and the control group auctions, we address this concern by excluding auctions that are more likely to face this problem. In particular, the top two panels of Table 7 replicate the earlier baseline estimates using two different subsets of control group auctions: in panel (a) we exclude all auctions held in central regions, and in panel (b) we exclude all auctions won by any firm that ever participated in one of the Firm's auctions. The estimates in both panels are similar to those in the baseline estimates in terms of sign, magnitude and significance.

The second set of robustness checks involves the awarding procedures used. As argued above, all the auctions considered share many characteristics, but there are nevertheless subtleties of the regulation that might affect outcomes. All procurements in the sample occur via sealed bid, price-only auctions. But across auctions, differences in both auction procedures and awarding methods exist. Auctions where a restricted set of bidders is invited to bid can be used under certain conditions, and indeed this method is used for 87 out of the 330 auctions held by the Firm.[32] Panel (c) reports estimates excluding these 87 auctions. The results are qualitatively identical to those in the baseline estimates. Regarding the awarding criterion, 42 out of the 330 auctions are awarded via modifications of the lowest price rule. All modifications entail penalizing discounts considered "too good to be true."[33] The awarding criterion is always specified in the call for tenders, so bidders knew these

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

[33]The Firm used the flexibility given to it by the Code to experiment with three alternatives to the lowest price rule. One method entailed awarding the contract to the contractor offering the discount closet to the average discount offered, increased by 20 percent. A second method entailed randomly deciding - after the bids were submitted - whether the criterion to be used was that of the highest discount or that of the discount closest to an average of the submitted discounts (either their simple average or a trim mean working as in the Average Bid Auction system described in Decarolis (2014)).

42 auctions were different and this might have altered their bidding. Indeed, when we exclude these 42 auctions from the sample we discover some changes relative to the baseline estimates: in the first three columns of panel (d), the sign of the coefficient switches from negative to positive, although it is still not statistically significant and is still relatively small in magnitude (about 3 percent of the reserve price). The source of this change is evident from the estimates in the last three columns: the magnitude of the β_2 estimate more than halves, declining from -15 percent to -6 percent. This suggests that part of the decline in winning discounts observed for the Firm's auctions after $t5+1$ is due to the presence of auctions held under awarding rules that soften price competition. In absolute terms, the magnitude of the β_1 and β_2 estimates is nearly identical, suggesting that the price increase observed after $t5+1$ cancels out the initial price saving observed right after $t1$. Overall, these results, as well as the overall lack of significance reported in the first three columns, indicate that the bolstering of performance did not come at the cost of major price increases.

[INSERT TABLE 7 APPROXIMATELY HERE]

A similar conclusion is derived from our third robustness check. In Table 8, we evaluate potential problems with inference by using alternative methods for standard errors. The four columns report 95 percent confidence interval estimates corresponding to models (2), (3), (5) and (6) of Table 6. The rows indicating "PA-Year" report estimates where the clustering is at the year and CA level, as in Table 6. The other rows in the table present two alternatives. The rows "CA" use clustering at the CA level only. As is well known from Bertrand, Duflo and Mullainathan (2004), this can serve to correct for overestimating the significance of the treatment effect driven by autocorrelation in the data. The table reveals that this correction has no qualitative implications for our results: relative to the baseline estimates, for all models involving both β_1 and β_2 there are no changes, while significance increases for models involving β_1 only. The latter models indicate a negative and significant effect at the 95 percent level. Since the clustering at CA and year is preferable to account for time variation, however, we prefer to rely on our more conservative baseline estimates. The second concern regarding inference is the fact that the level at which the treatment effectively takes place is that of the procurer and we observe exclusively one procurer, the

Firm, receiving the treatment. Hence, any shock hitting the Firm at $t1$ biases the estimate of β_1. As argued by Conley and Taber (2011), if the shocks potentially hitting the Firm and the control CAs belong to the same distribution, and if a sufficiently large number of control CAs are observed, valid inference can still be conducted by adjusting the standard errors. Since we have many control CAs, we use their method to asses how significance changes relative to our baseline estimates. The "Conley-Taber" rows are indeed different from the baseline ones: in columns (3) and (4), β_1 loses significance, while β_2 loses significance in model (4) when the largest set of control CAs is used. Overall, this indicates that we should be cautious in interpreting the findings in Table 6 about significant and opposite signs of β_1 and β_2. Hence, a more conservative interpretation is that there are no statistically significant price changes throughout the sample.

[INSERT TABLE 8 APPROXIMATELY HERE]

VI Discussion: Adverse Selection and Moral Hazard

The three motives offered by the literature to understand why competitive procedures such as first price auctions induce a trade-off between price and performance are adverse selection, moral hazard and the winner's curse fallacy. The latter refers to bidders' inability to properly assess project costs at the time of bidding. In our setting, this non-equilibrium phenomenon is unlikely to play a major role, as we observe experienced contractors repeatedly bidding in auctions for relatively simple contracts. It is interesting, however, to understand the extent to which selection and moral hazard can explain our findings, as they can have different implications for how best to design systems to integrate past performance in procurement markets.[34]

The Firm's experiment might have induced responses in both of them: performance improvements can derive from either more effort in the execution stage by contractors, or

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

by a better selection of contractors, or by a combination of both. Indeed, Figure 10 presents evidence consistent with both by showing how the cdf of winning bids in the Firm's auctions evolves between those held before $t1$, after $t5+1$, and in between these two periods. The noteworthy aspect is the disappearance post $t1$ of right tail discounts, representing discounts of one third or more relative to the reserve price. It is precisely this type of abnormally high discounts that procurers worry will be associated with poor performance. Since replicating Figure 10 for those firms bidding both before and after $t1$ leads to a similar finding of a disappearing right tail after $t1$, we can conclude that the altered bidding behavior of these suppliers is compatible with the presence of moral hazard in contractual performance.

[INSERT FIGURE 10 APPROXIMATELY HERE]

More specifically, it is useful to describe how a stylized model of a first price sealed bid procurement auction with moral hazard can rationalize our findings. Equilibrium bids should depend on two elements: production costs, $C(e)$, which are an increasing function of the effort e that the bidder plans to put in place in the execution stage; and a strategic markup, $M(n)$, which is inversely proportional to the number of competitors, n. Prior to $t1$, each auction exists in isolation - the outcome of an auction does not matter for future auctions.[35] Thus, bidders will choose low effort levels to reduce their cost and maximize their profits. From $t1$ onward, however, even if the awarding rule remains the lowest price, the game played by the contractors becomes dynamic: winning an auction can imply being audited and, hence, an opportunity to modify one's own RI while reducing the rival's chances of improving their RI. This likely implies changes to both components of the bid relative to the pre-$t1$ case: if better compliance requires more effort, then the optimal $C(e)$ will likely be higher. Moreover, the strategic markup now depends not only on n, but also on the distribution of RI across bidders.[36] Finally, the bid now also incorporates a third

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

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

element: the continuation value associated with the evolution of the RI. Indeed, winning today and earning a good RI is expected to produce savings in the stream of future auctions, once the scoring rule auction is introduced. This continuation value increases the value of winning today and, hence, balances increases in production costs. Clearly, the relevance of the continuation value depends on how many auctions suppliers perceive they will be able to use their good RI for.

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

Thus, an explanation for the patterns observed in the data is that, right after $t1$, the increase in $C(e)$ was dominated by the changes in the strategic markup and the continuation value. After contractors accumulated a good RI, however, the value of winning an auction in the pre scoring rule period declines as obtaining positive audit reports cannot offer a competitive edge over rivals. Thus, in this phase the increased production cost dominates.[37]

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

In the data, however, the effects of the new system concerned not only bidding and performance, but also participation choices. Indeed, while the summary statistics show that the number of bids submitted remains stable and approximately equal to 10 both before and after $t1$, the set of bidders changed in the Firm's auctions: while there are 34 suppliers placing at least one bid both before and after $t1$, there are other 36 suppliers who place at least one bid before $t1$, but no bid afterwards. We refer to the latter group of firms as "*exiters*" and to the former as "*stayers*." There are also 3 new entrants placing bids only after $t1$, but never before then. This implies that the average number of bids placed per bidder doubles: from 0.14 (i.e, 10/70) to 0.27 (i.e., 10/37). This increased participation is due to the *stayers*, not to unusually high bidding frequencies for the 3 new entrants. It is likely driven by the same incentive to earn RI that we discussed when analyzing the evidence on winning bids. As regards *exiters* and new entrants, however, their mere presence potentially indicates that the experiment might have also triggered some selection effects.

If we focus on *exiters*, however, the data provides only weak evidence of possible selection effects.[38] In particular, Figure 11 shows the timing of the exits does not seem clearly linked to $t1$. This figure reports the last date at which each of the *exiters* (represented by the numerical identifiers on the vertical axis) placed a bid. The smooth path of exits indicates more of a gradual process than a sharp drop at $t1$.

[INSERT FIGURE 11 APPROXIMATELY HERE]

Furthermore, as illustrated by Figure 12, if we compare the cdf of winning bids by both *exiters* and *stayers* (in the pre-$t1$ auctions), we do not observe significant differences. Finally, even in terms of characteristics, *exiters* do not seem to be substantially different from *stayers*. Table 12 reports summary statistics for the subset of *exiters* and *satyers* that we could match to the Infocamere database, the Italian firm registry. Statistics for the *exiters* are reported in the first four columns of panel (a), followed by statistics for the *stayers* in the following four columns. Along most dimensions, *exiters* are smaller than *stayers*; this is the case for revenues, profits and capital. The average number of employees is also

[38]For the 3 new entrants, the type of analysis performed below for the *exiters* cannot be replicated as only one of them could be matched to the Infocamere database.

lower, but in this case the median is nearly identical. For both groups, the wide variation in characteristics among firms means that the differences in the averages are not statistically significant and it is not obvious to how to interpret the results. Thus, to benchmark these statistics we present in panel (b) the analogous statistics obtained for the suppliers active in the auctions of the multi-utility company of the city of Turin. This is the multi-utility company that awards most contracts within the DD control group. Analogously to what was done for the Firm, we partition its suppliers into those bidding both before and after $t1$ (*stayers*) and those bidding only before $t1$ (*exiters*). The comparison of the two groups leads to similar conclusions to those found for the Firm's suppliers: the average revenues, profits and capitals are higher among *stayers*. But the data are again characterized by many extreme observations and the result is reversed for revenues and profits when looking at the median.

[INSERT FIGURE 12 APPROXIMATELY HERE]

[INSERT TABLE 9 APPROXIMATELY HERE]

We conclude that, overall, there is no strong evidence that the pool of *exiters* in the Firm's auctions is selected in any particular way relative to the typical exit behavior in the market. Thus, the effects that we uncovered in the earlier section are likely driven to a large extent by changes in the behavior of the Firm's contractors.

VII Conclusions

This paper has studied the merits of using past performance to spur greater efforts from contractors when executing public works. The evaluation of the evidence from an experiment undertaken by a large Italian multi-utility company has shown strong improvements in performance after the firm announced the future use of the past performance scores to award future contracts. To some extent this may resemble the well-known *Hawthorne effect*. However, contrary to the *Hawthorne effect*, the improvement was not short-lived, even

if we consider that the contractors could have stopped trusting the firm over the delayed implementation of the new awarding rule, and that it was easier for contractors to improve their score when the starting point was lower than later, when the marginal cost to improve became higher. Regarding prices, we find some evidence of a moderate increase in prices after suppliers have achieved high scores for their performances. However, the overall price increase appears rather small when compared to the substantial improvement in performance.

Although the empirical evidence in this paper is suggestive of the benefits from implementing reputation mechanisms in public procurement, an extension of our analysis would involve exploring welfare implications. In particular, the firm's managers who designed the experiment had among their main goals improving safety to the point of eliminating any serious accidents to the workers involved in the contracts. Their reference was Heinrich's "pyramid," a statistical relationship that in the context of industrial systems is used to argue that, on average, for every 1,200 deviations on small tasks there will be 600 quasi-accidents, 30 incidences of material damage, 10 minor accidents and 1 major accident. It would be therefore interesting to assess the extent to which the improved performance obtained among the firm's contractors translated into fewer accidents.

Furthermore, although several different mechanisms might explain why the increased quality and safety achieved was not reflected into substantially higher prices, it is interesting to note that the explanation offered by the management of the firm is that most of the gains came from improvements in management practices within contractors. Thanks to new data on management practices collected in the last ten years through the World Management Survey,[39] there has been increased attention on the role of management in explaining productivity differences (Bloom et al. (2014)). In this respect, exploring the details of the managerial changes implemented by the suppliers would be useful to understand how (the announcement of) new procurement rules triggered an improvement in management. Regarding this, it is also important to highlight that, while we have stressed the public procurement implications of our analysis, our findings are also relevant to private procurement

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

practices, where the use of vendor rating systems is widespread but little is know about their effectiveness.

Finally, once the merits of this kind of reputation mechanism in improving contractor performance are proven, many aspects remain open and give room for future research. For example, how to optimize the weigh in the scoring rules in different sectors, how to discipline the rating for new entrants, how to structure the weights in the awarding criteria, and how to choose the optimal "memory" of the indicator (i.e. how long the window of time over which the RI is calculated should be and how heavily older information should be discounted).

References

Albano, Gian Luigi, Berardino Cesi, and Alberto Iozzi. 2011. "Relational Procurement Contracts: A Simple Model of Reputation Mechanism." *SSRN WP*, 1884979.

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

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

Banerjee, Abhijit V., and Esther Duflo. 2000. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics*, 122(3): 1235–1264.

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

Beuve, Jean, and Lisa Chever. 2014. "Quality of Outsourced Services, Opportunism and Contract Design. Evidence from Cleaning Contracts." *mimeo*.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, Daniela Scur, and John Van Reenen. 2014. "The New Empirical Economics of Management." National Bureau of Economic Research, Inc NBER Working Papers 20102.

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

Bulow, Jeremy, and Paul Klemperer. 1996. "Auctions Versus Negotiations." *American Economic Review*, 86(1): 180–194.

Bulow, Jeremy, and Paul Klemperer. 2009. "Why Do Sellers (Usually) Prefer Auctions?" *American Economic Review*, 99(4): 1544–1575.

Butler, Jeffrey V., Enrica Carbone, Pierluigi Conzo, and Giancarlo Spagnolo. 2014. "Reputation and Entry in Procurement." *Working paper*.

Cabral, Lus M. B., and Miguel Villas-Boas. 2005. "Bertrand Supertraps." *Management Science*, 51(4): 599613.

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

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

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

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

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

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

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

Gil, Ricard, and Justin Marion. 2012. "Self-Enforcing Agreements and Relational Contracting: Evidence from California Highway Procurement." *Journal of Law, Economics, and Organization.*

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

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

Kim, In-Gyu. 1998. "A model of selective tendering: Does bidding competition deter opportunism by contractors?" *The Quarterly Review of Economics and Finance*, 38(4): 907–925.

Klein, Tobias J., Christian Lambertz, and Konrad O. Stahl. 2016. "Market Transparency, Adverse Selection, and Moral Hazard." *Journal of Political Economy*, Forthcoming.

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

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

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

Lewis, Gregory, and Patrick Bajari. 2013. "Moral Hazard, Incentive Contracts, and Risk: Evidence from Procurement." *The Review of Economic Studies*.

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

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

Manuel, Kate M. 2015. "Evaluating the "Past Performance" of Federal Contractors: Legal Requirements and Issues." *Congressional Research Service*, R41562.

Marion, Justin. 2016. "Affirmative Action Exemptions and Capacity Constrained Firms." *Mimeo*.

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

Saussier, Stephane, and Jean Tirole. 2015. "Strengthening the Efficiency of Public Procurement." *Les notes du conseil d'analyse economique*, 22.

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.

Tables and Figures

Table 1: Reputation Index Components

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

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

Table 2: Summary Statistics - Audit Data

Class	Category	Number of observations	PI Share	Share Compliant Pre t1	t1-5	Post t5
Safety						
	Equipment and machinery	6,761	.238	.701	.899	.942
	Documentation	9,407	.212	.333	.541	.721
	Works execution	6,311	.231	.187	.753	.88
	Personnel	3,875	.228	.319	.602	.684
	Works site regularity	13,615	.204	.103	.449	.686
	Works site safety	14,730	.231	.306	.614	.813
	H.T. works site controls	351	0	-	.62	.801
Quality						
	Works on joints	1,471	.206	1	.939	.976
	Customer relationship mgnt	69	0	1	.75	1
	Air works	107	.009	-	.963	1
	Underground works	7,607	.112	.398	.634	.715
	Works on transformer station	233	0	1	.993	1
Sample size:		64,537	13,274	1,997	31,478	31,062

This table reports summary statistics for the Audits data. The 64,537 scores assigned to each parameter audited were produced through 1,951 individual audits that took place on 187 distinct contracts.

Table 3: Summary Statistics - Auctions Cross Section

*Pre t*1

	(1) Treated			(2) Control		
	Mean	SD	N	Mean	SD	N
Winning Discount	22.84	10.45	206	21.39	10.17	1568
Winning Bid	730.7	550.8	206	451.0	524.9	1568
Durata	387.5	170.9	167	329.0	335.8	1406
Num. Bids	9.870	5.668	154	.	.	0
Public Illumination	0.223	0.417	206	0.249	0.433	1568
Central Region	1	0	206	0.214	0.410	1568
Municipal Firm	1	0	206	0.430	0.495	1568

*Post t*1

	(1) Treated			(2) Control		
	Mean	SD	N	Mean	SD	N
Winning Discount	20.59	10.38	124	22.73	11.85	1715
Winning Bid	570.4	402.2	124	417.5	490.0	1715
Durata	346.0	103.7	21	354.8	1204.9	1409
Num. Bids	10.60	4.836	104	.	.	0
Public Illumination	0.282	0.452	124	0.241	0.428	1715
Central Region	1	0	124	0.171	0.377	1715
Municipal Firm	1	0	124	0.434	0.496	1715

This table reports summary statistics for the Auctions data. The control group observations are from the union of Control Group 2 and 3.

Table 4: Chow and Bai-Perron Tests

(a) Chow Tests

	Weighted Compliance		*Quality*		*Safety*	
	1 break at t1	5 breaks at t1-5	1 break at t1	5 breaks at t1-5	1 break at t1	5 breaks at t1-5
F-statistic	27.80	117.20	68.48	21.77	25.61	148.84
	(0.0000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

(b) Bai-Perron Tests

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

The table reports the results of Chow (top panel) and Bai-Perron (bottom panel) tests. The variable is the monthly weighted average compliance, measured on all audited parameters (first two columns) or on the subset of quality parameters (next two columns) or safety parameters (latter two columns). For the Bai-Perron test, the criterion used is the that of sequential F-statistic determined breaks.Analogous results, however, are obtained when using the criterion of the significant F-statistic largest breaks.

Table 5: Probability of Compliant Parameter

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

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

Table 6: Baseline Estimates

Panel (a): All Contracting Authorities						
	(1)	(2)	(3)	(4)	(5)	(6)
β_1	-3.93	-3.99	-3.90	5.69***	5.63***	5.28***
	(4.49)	(4.49)	(4.27)	(0.97)	(0.98)	(1.08)
β_2				-14.69***	-14.70***	-14.02***
				(3.44)	(3.44)	(3.28)
N	3613	3613	3613	3613	3613	3613
R^2	0.41	0.41	0.42	0.42	0.43	0.43

Panel (b): Contracting Authorities in Central Regions						
	(1)	(2)	(3)	(4)	(5)	(6)
β_1	-4.91	-4.91	-4.97	6.18***	6.18***	5.33***
	(4.19)	(4.19)	(3.86)	(1.26)	(1.26)	(1.47)
β_2				-17.47***	-17.47***	-16.17***
				(3.05)	(3.05)	(2.84)
N	959	959	959	959	959	959
R^2	0.21	0.21	0.23	0.27	0.27	0.28
CA-type FE	No	Yes	Yes	No	Yes	Yes
Contract Chars FE	No	No	Yes	No	No	Yes

This table contains the baseline DD estimates. Standard errors clusters by year and CA are reported in parenthesis. The two panels differ for the control group used: panel (a) uses all available auctions, while panel (b) uses only auctions held by contracting authorities in central regions. In each panel, the first three columns report estimates for the model in equation (4), while the last three columns report estimates for the model in equation (5). For each model, estimates for three specifications differing on the set of covariates, X, are presented: we first include in X only the constant, then add procurer characteristics and, finally, also add contract characteristics.

Table 7: Robustness Checks: Contamination and Sample Effects

Panel (a): No Contracting Authorities in Central Regions						
	(1)	(2)	(3)	(4)	(5)	(6)
β_1	-3.64	-3.69	-3.55	5.63***	5.57***	5.26***
	(4.27)	(4.27)	(4.04)	(0.73)	(0.73)	(0.82)
β_2				-14.19***	-14.18***	-13.50***
				(3.35)	(3.35)	(3.21)
N	2984	2984	2984	2984	2984	2984
R^2	0.44	0.44	0.45	0.45	0.45	0.46
Panel (b): No Auctions with Common Winners						
	(1)	(2)	(3)	(4)	(5)	(6)
β_1	-4.06	-4.13	-4.01	5.69***	5.63***	5.33***
	(4.51)	(4.51)	(4.31)	(1.00)	(1.02)	(1.14)
β_2				-14.79***	-14.80***	-14.15***
				(3.47)	(3.46)	(3.31)
N	3585	3585	3585	3585	3585	3585
R^2	0.41	0.42	0.42	0.43	0.43	0.43
Panel (c): No Restrictions to Open Competition						
	(1)	(2)	(3)	(4)	(5)	(6)
β_1	-4.13	-4.19	-3.76	6.33***	6.27***	6.30***
	(4.42)	(4.42)	(4.26)	(0.68)	(0.68)	(0.85)
β_2				-15.11***	-15.11***	-14.60***
				(3.45)	(3.44)	(3.30)
N	3526	3526	3526	3526	3526	3526
R^2	0.42	0.42	0.43	0.43	0.43	0.44
Panel (d): No Variations to the Lowest Price Criterion						
	(1)	(2)	(3)	(4)	(5)	(6)
β_1	3.70	3.62	3.43	6.52***	6.46***	6.02***
	(2.51)	(2.53)	(2.38)	(1.36)	(1.38)	(1.39)
β_2				-6.02***	-6.04***	-5.50***
				(0.78)	(0.78)	(0.80)
N	3531	3531	3531	3531	3531	3531
R^2	0.43	0.43	0.43	0.43	0.43	0.44
CA-type FE	No	Yes	Yes	No	Yes	Yes
Contract Chars FE	No	No	Yes	No	No	Yes

This table contains results to evaluate the robustness of the baseline DD estimates in Table 6 with respect to control group contamination (top two panels) and features of the awarding methods (bottom two panels).

Table 8: Robustness: Inference

Panel (a): All Contracting Authorities

	(1)	(2)	(3)	(4)
VARIABLES	W.Discount	W.Discount	W.Discount	W.Discount
CA-Year	(-8.4;4.7)	(-8.3;4.7)	(2.5;5.2)	(2.4;5.3)
CA	(-2.9;-0.8)	(-2.9;-0.7)	(2.9;4.8)	(2.8;4.8)
Conley-Taber	(-6.7;6.2)	(-6.4;6.2)	(-3.5;9.6)	(-3.1;9.6)
CA-Year			(-14.3;-6.7)	(-14.0;-6.8)
CA			(-12.1;-9.0)	(-11.9;-8.8)
Conley-Taber			(-31.7;-0.1)	(-31.0;0.3)

Panel (b): Contracting Authorities in Central Regions

	(1)	(2)	(3)	(4)
VARIABLES	W.Discount	W.Discount	W.Discount	W.Discount
CA-Year	(-10.2;4.5)	(-10.1;4.1)	(1.7;6.7)	(1.0;6.6)
CA	(-5.2;-0.5)	(-5.4;-0.6)	(2.2;6.2)	(1.9;5.7)
Conley-Taber	(-7.4;0.3)	(-7.0;-0.5)	(-2.8;4.7)	(-2.5;3.9)
CA-Year			(-17.8;-8.2)	(-17.0;-8.2)
CA			(-16.8;-9.3)	(-16.4;-8.8)
Conley-Taber			(-13.3;-5.9)	(-13.0;-6.5)

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

Table 9: Summary stats: Exiting and Incumbent firms

Panel (a): Contractors Entering the Firm's Auctions

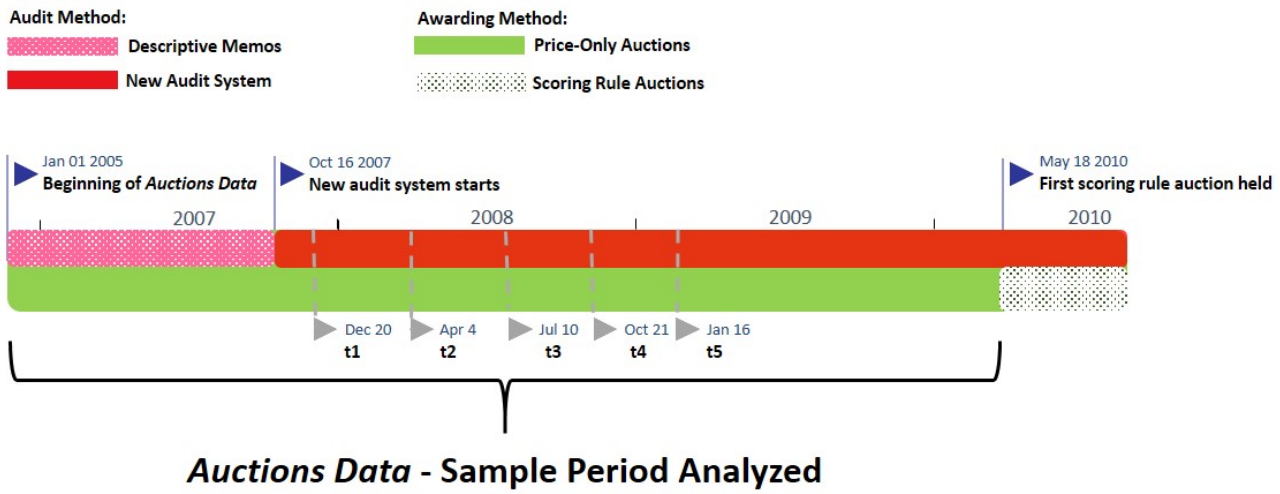
	Exiters				*Stayers*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean	p50	SD	N	Mean	p50	SD	N
Revenues	8,283	2,458	14,615	24	8,934	5,660	9,401	16
Profits	-21	6	697	24	32	5	73	16
Capital	391	36	788	24	998	47	2699	16
Number of Employees	10.3	5	11.1	24	51.7	4.50	180.4	16
Number of Managers	4.96	2	7.57	24	3.38	2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

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

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

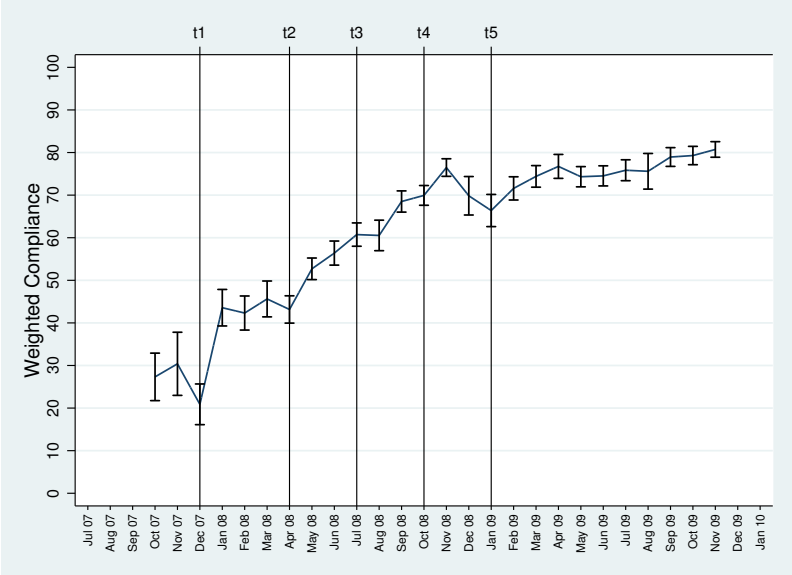
The table reports summary statistics for the contractors bidding in the auctions. Panel (a) refers to the contractors active in the Firm's auctions, while panel (b) refers to the contractors bidding in the auctions of Turin's multi-utility company (IRIDE). Across all multi utilities in the DD control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

Figure 1: Timeline



The chart illustrates the time span of the data and the timing of the Firm's announcements. The *Auctions* dataset covers the period between 1/1/2005 and 4/1/2010. All auctions held in this period are price-only auctions. The five Firm's announcements, marked $t1$,...,$t5$, were held after the new auditing system started to inform suppliers about the functioning and the intended usage of this system. The implementation of the awarding rule based on both price and past performance is outside the analysis sample.

Figure 2: Average Compliance

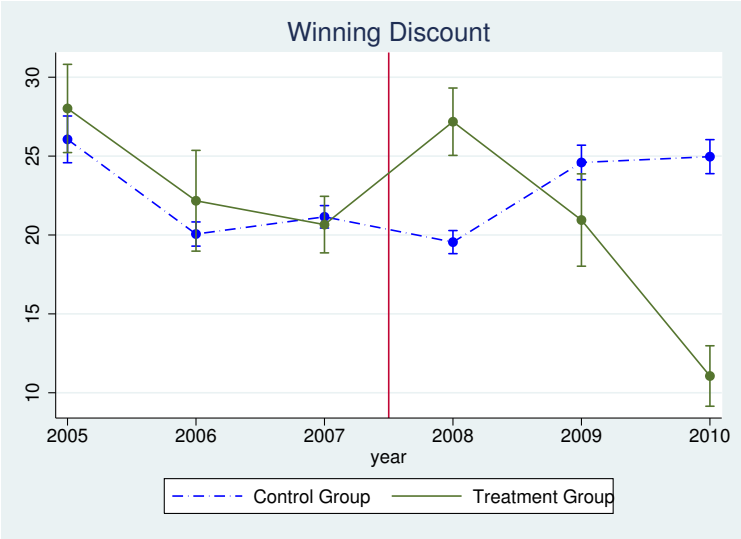


Source: Audits data. The line shows the monthly average compliance calculated on all parameters inspected in the month of reference, weighting each parameter by its weight in the RI. The vertical lines identify each announcement date.

Figure 3: Winning Discounts (Dec 2004 - May 2010)

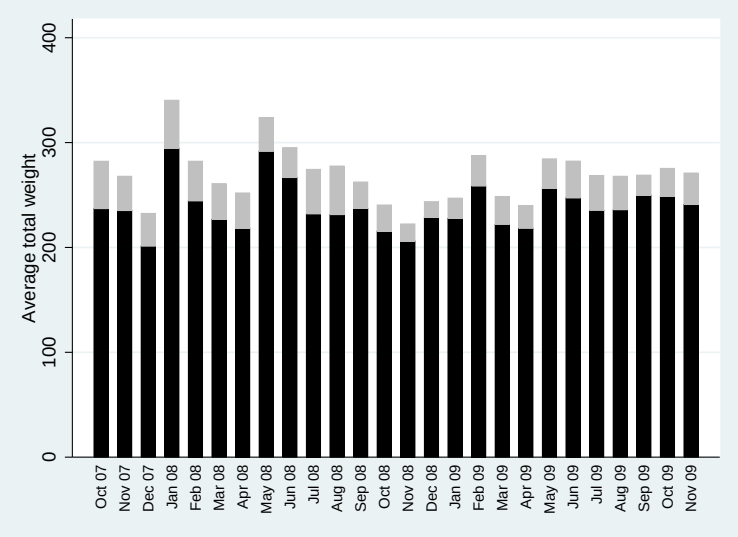
Source: Auctions data. The scatter plot reports all the winning discounts placed in the Firm's auction during the sample period. Note that the sample period is longer relative tot hat of the Audits data.

Figure 4: Evolution of Winning Bids



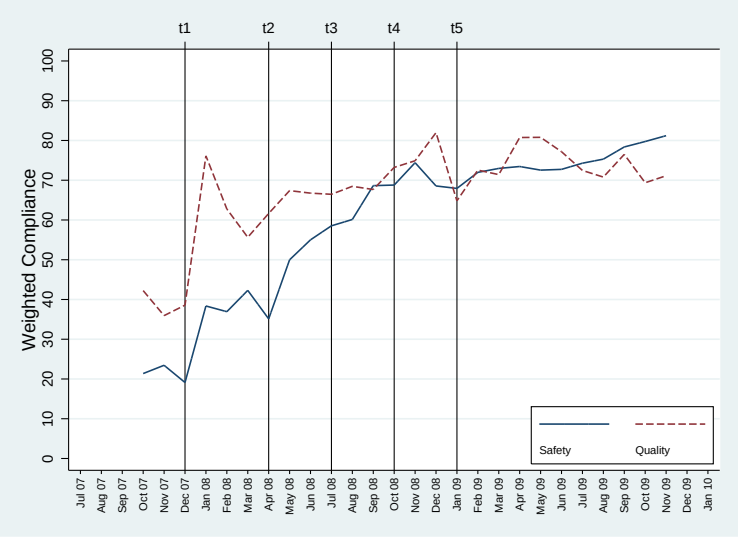
Sample: Auctions data. Evolution over time of the average winning discount for both treated (the Firm) and control (other CA) procurers.

Figure 5: Safety and Quality: Average Weights across Audits



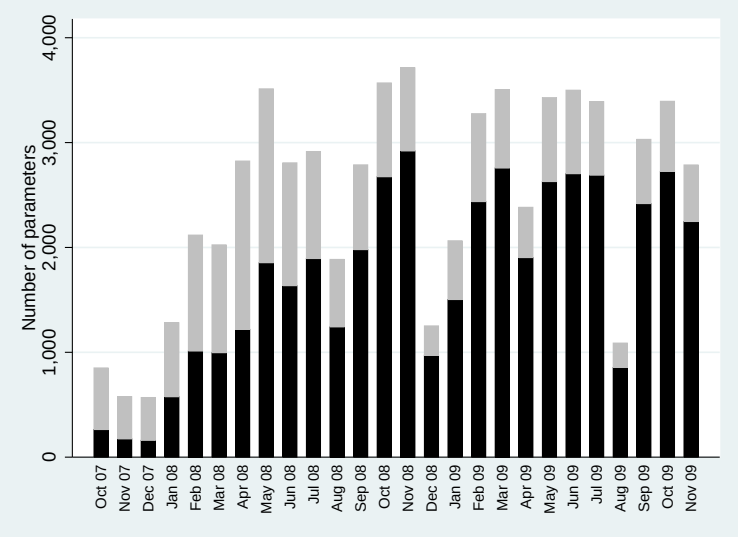
Source: Audits data. The plot represents the total weight, by audit, of parameters relating to Quality dimension (grey bar) and Safety dimension (black bar).

Figure 6: Safety and Quality: Evolution of Compliance over Time



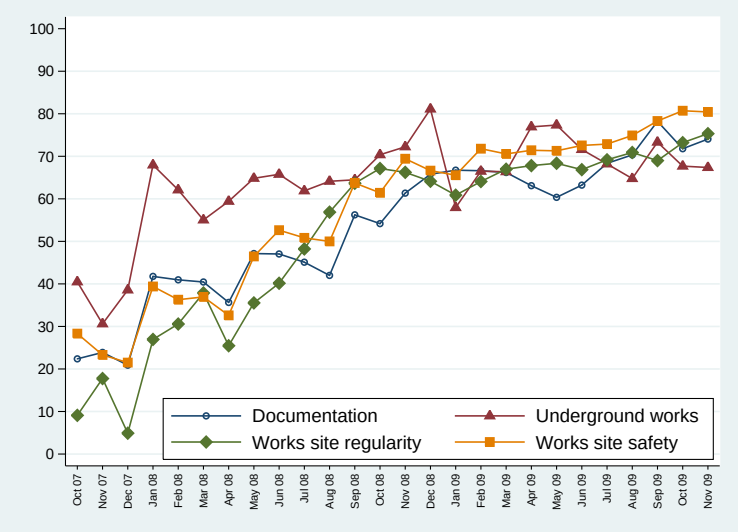
Source: Audits data. Monthly average compliance calculated separately for Safety and Quality on all parameters inspected in the month of reference, weighting each parameter by its weight in the RI. The vertical lines identify each announcement date.

Figure 7: Parameters Audited: Frequency of Checks



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

Figure 8: Parameters Audited: Evolution of Compliance over Time

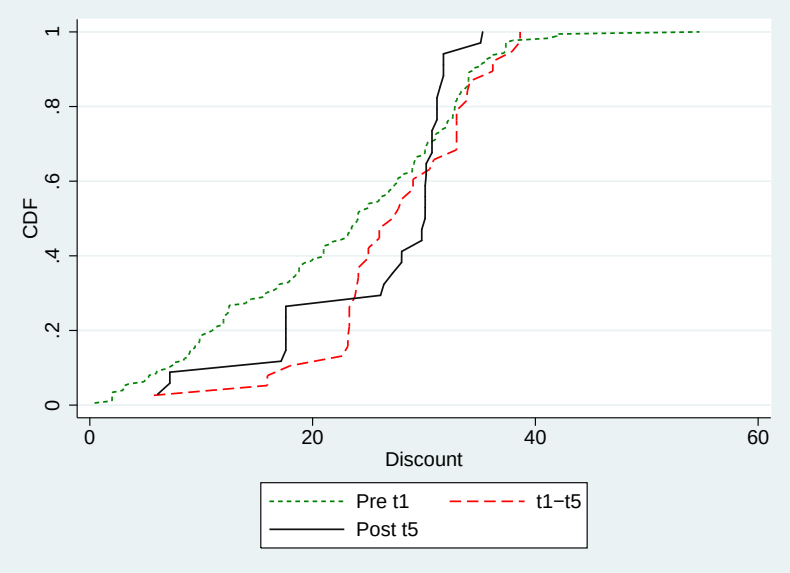


Source: Audits data. The plot shows the progress of the reputation index calculated on a monthly basis for each of the four most audited Safety and Quality dimension.

Figure 9: Contractors: Evolution of Compliance over Time

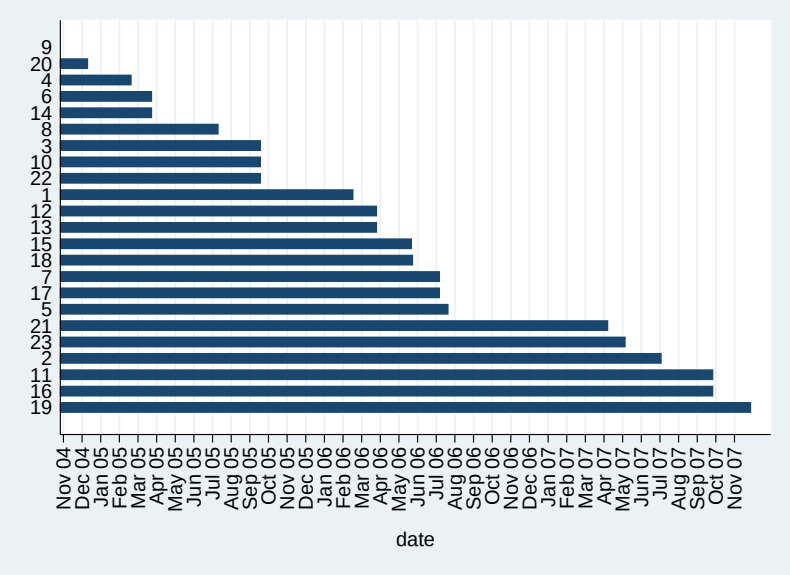
Source: Audits data. The four lines show the progress of the reputation index, calculated on a monthly basis, for 4 different groups of firms. The groups are formed on the basis of the firm's successfulness in concluding contracts. The line with circle markers represent the "most awarded" firms, the triangle is for the "often awarded" group, the diamond is for "the less awarded group" and the square is for "the rarely awarded group."

Figure 10: Discount CDF Pre-t1, t1-to-t5 and Post-t5



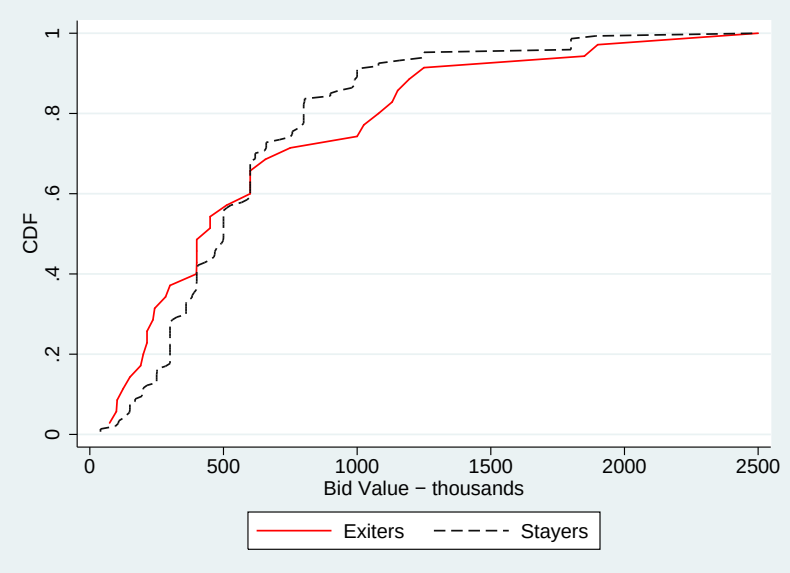
Source: Auctions data. The plot represents the cdf of the winning bid, dividing binds depending on the timing of the auction relative to $t1$ and $t5 + 1$.

Figure 11: Last auction date participated (*exiters*)



Source: Auctions data. Each bar represents the time until when the supplier last bids. The figure is drawn for the sample of *exiters* only.

Figure 12: Bid CDF for exiting and incumbent firms



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