Can the Private Sector Ensure the Public Interest? Evidence from Federal Procurement*

Leonardo M. Giuffrida and Gabriele Rovigatti†

December 3, 2018

Abstract

We empirically investigate the effect of procurement oversight on contract outcomes. In particular, we stress a distinction between *public* and *private* oversight: the former is a set of bureaucratic checks enacted by contracting offices, while the latter is carried out by private insurance companies whose money is at stake through the so-called performance bonding. By focusing on the U.S. federal service contracts in the period 2005-2015, we exploit an exogenous variation in the threshold for the application of both sources of oversight in order to separately estimate their causal effects on execution costs and time. We find that: (i) private oversight has a positive effect on outcomes through the screening of bidders that alters the pool of winning firms; (ii) public oversight negatively affects outcomes, due to excessive red tape induced by low-competence buyers.

JEL: D73, D82, H44, H57.

Keywords: oversight, procurement, screening, red tape, moral hazard.

*Special thanks go to Francesco Decarolis, without whom this work would not have seen the light of day. We are grateful to Jean Beuve, Stefano Gagliarducci, Juan-José Ganuza, Sofia Lundberg, Calvin Luscombe, Stéphane Saussier, Giancarlo Spagnolo, and Hidenori Takahashi, who provided crucial insights and expertise. Also, we thank Klenio Barbosa, Florian Buhlmann, Decio Coviello, Ruben Enikolopov, John Gatherhood, Alessandro Gavazza, Alberto Iozzi, Eckhard Janeba, Vitalijs Jascisens, Neale Mahoney, Juan Ortner, Johannes Spinnewijn, Francis Vella, Han Ye, eight anonymous referees as well as the audiences at the Tor Vergata Lunch Seminars, Barcelona Jamboree 2016, GRAPE February 2017 meeting, EARIE 2017, WEEE 2017, SIEP 2017, UniBz Applied Microeconomics Workshop 2018, JEI 2018 and University of Mannheim Research Seminars in Public Economics for comments that improved the manuscript. The authors also thank the Società Italiana di Economia Pubblica for having awarded this paper with the SIEP prize at the 29th meeting in September 2017. Any errors are our own.

†Giuffrida: ZEW Mannheim. Mail: *giuffrida@zew.de*. Rovigatti: Bank of Italy. Mail: *gabriele.rovigatti@esterni.bancaditalia.it*

I Introduction

Efficient contract procurement is a complex task. Sellers have private information on production costs that are not fully disclosed through bids, and incentives in exerting sub-optimal levels of effort once awarded the contract. These sources of asymmetric information between buyer and seller, combined with the intrinsic cost uncertainty at the awarding stage, paves the way for the emergence of adverse selection and moral hazard during the procurement process (Laffont and Tirole, 1990; Bajari and Lewis, 2014). In turn, these issues jointly lead to the renegotiation of contract terms, increases in costs and time to completion and, ultimately, efficiency losses. Public procurement accounts for 14 percent of GDP in OECD member countries and it draws its budget on public resources.[1] Dealing with the above frictions is a first-order concern for public procurers as contract inefficiency is at tax payers expense.

To cope with this well-known phenomenon, the empirical literature so far has separately centred on the awarding phase and the operational phase. On the one hand, contributions mostly focused on the former, that is on role of awarding procedures in screening bidders and optimal contract design in avoiding misbehavior.[2] On the other hand, albeit in the practice of public procurement the handling of the contract execution stage is also seen as a first-order concern, a few contributions (Bajari, Houghton and Tadelis, 2014; Lewis and Bajari, 2011) looked at the operational phase. Yet, an efficient procurement regulation should consider both sides of the coin and require a balanced level of global contract management tools - i.e., tools aimed at alleviating both adverse selection before contract award and moral hazard during project execution. Throughout the paper, we label such tools "oversight". Although the optimal level of oversight is well defined in theoretical contributions (Shavell, 1984), however it is fiercely debated in practical applications.[3] This work aims at filling this gap by providing empirical evidence that oversight in public procurement matters.

We propose a distinction between *public* and *private* oversight, depending on its source.

[1]Source: http://www.oecd.org/governance/public-procurement/ accessed on May 31, 2018.

[2]Bajari, McMillan and Tadelis (2009) and Decarolis (2014) belong to the former group, Bajari and Tadelis (2001) to the latter.

[3]See for example the technical reports GAO (2013) and Garvin et al. (2011).

Public oversight includes all formal checks - solicitation procedures, cost certifications, pricing data transmission, production surveillance - which the contracting authorities enact during both the contract awarding process and the project execution. It typically involves considerable paperwork for buyer and sellers and takes discretionary power away from contracting officers but, at the cost of some *red tape*, it is aimed at mitigating adverse selection and moral hazard (Kaufman, 1977; Shleifer and Vishny, 1998). Combinations of red tape and bureaucrats' ineptitude might be deadly for procurement processes: in February, 2014, the U.S. Court of Appeals for the Federal Circuit sentenced against the federal government because the U.S. Navy officials took two years after the original completion date to accept the project as complete and caused million dollars of losses to the contractor.[4] By contrast, private oversight involves third parties - surety companies - issuing bonds (*performance bonds*) to secure the buyer against unpredictable events.[5] If the seller fails to fulfill contractual tasks, contracting authorities make claims to recover losses and the underwriting surety is called upon either to complete the project by itself (i.e., with its own resources or by subcontracting) or, as last instance, to refund the authority of the bond value.[6] Being liable in case of unsatisfactory contract outcomes provides a strong incentive for sureties to both screen bidders (ex ante) and monitor contractors (ex post). They help relieve the asymmetry of information between the buyer and the seller through the screening enacted by price discrimination on premia, which directly affects the bids placed by potential contractors.[7] Hence, private oversight enhances the selection of the best contractors and provides a second tier of monitoring of contractors' progresses.

Identifying the extent to and the channels through which public and private oversight affect contract outcomes has clear policy implications. Moreover, the performance bonding is an increasingly popular tool in procurement governance and in many countries there is an ongoing debate about their efficacy; this paper contributes by providing quantitative

[4] *Metcalf Construction Company, Inc. v. United States* began a control for disputes with the federal government and also provides rationale useful to contractors in disputes with any public or private project owner. See http://caselaw.findlaw.com/us-federal-circuit/1656993.html.

[5] Surety companies (more simply *sureties*) usually are subsidiaries of insurance companies.

[6] The rationale of the law was initially to protect the buyer from losses in case of seller's bankruptcy.

[7] Premia paid for performance bonds may vary depending on the valuation of bidder quality. Hence, *ceteris paribus*, worse contractors face higher premia, their bids are higher and their likelyhood to win the auction lower.

support.[8] The U.S. constitutes an excellent case study for the outlined framework as both public and private oversight are required depending on the industry and contract value. Furthermore, performance bonding is well-known among all players in the procurement market in the U.S. as it was the first country to introduce it in 1894.[9]

In this paper we use a recently available database containing contract-level information on the universe of U.S. federal purchases.[10] Focussing on 2005-2015 procurements, our identification strategy relies on the contemporaneous change (occurring on October 1, 2010, that is the beginning of fiscal year 2011) of the threshold for (i) the Simplified Acquisition Procedures, exempting *all* federal procurement contracts from public oversight; and (ii) the Miller Act, the law requiring private oversight *only* in construction projects through performance bonding.[11] We use a difference-in-difference-in-differences (DDD) approach to estimate the causal effect of the different sources of oversight on performance outcomes.[12] Specifically, for the whole population of federal service procurement contracts, we compare the average change in outcomes of contracts that are exempted from public oversight with corresponding changes among those that remain subject to the requirement. To take into account possible differences between constructions and other services, due to the additional application of private oversight to the former procurement category, we simultaneously compare changes in outcomes between the two groups.

Backed up by a battery of robustness checks on sample selection, suitability of empirical strategy and risk of differential shocks, our reduced-form analysis yields two main findings. First, exempting contracts from private oversight negatively affects performance in terms of time and cost, worsening it by 9 and 4.2 percent, respectively. We exploit firm-level data to

[8] Performance bonds are widely used at government-level procurement not only in the U.S., but also in Japan and Canada. Also, many states in the U.S. have introduced performance bonding through the so-called "Little Miller Act". In 1999, the European Commission's Enterprise Section published a report titled "Abnormally low tenders" with detection and rejection rules for abnormally low tenders and started a working group on performance bonds (European Commission Enterprise Section, 1999).

[9] The *Heard Act*, requiring performance bonds on all federally funded projects, was replaced by the *Miller Act* in 1935.

[10] The Federal Procurement Data System - Next Generation (FPDS-NG) is publicly available at https://www.fpds.gov/fpdsng_cms/index.php/en/ and updated on a daily basis. The FPDS database is well documented and was recently used by Liebman and Mahoney (2017), among the others.

[11] The subset of service contracts totals around $5.6 trillion in government expenditure.

[12] Recently, Bergman et al. (2016) used the same econometric approach in the procurement of elderly care services in Sweden.

provide evidence that adverse selection plays a key role in driving our estimates of private oversight. All the above findings are in line with Calveras, Ganuza and Hauk (2004), who develop a model of public procurement with performance bonding, where the premium paid is proportional to the riskiness of the bidder, and show that the presence of private oversight improves the selection of winning firms.[13] Second, we find that exempting contracts from public oversight improves both time and cost outcomes, leading to increases in performance of 7.2 percent and 5.3 percent, respectively. This is in line with the results of Calvo, Cui and Serpa (2016). In addition, we find that the red tape effect in public oversight is negatively correlated with the contracting authority quality, measured in terms of past performances, and we do not find any significant outcome when estimating the treatment effect on the subset of high-competence offices.[14] In the construction sector, where the 2011 reform implied the simultaneous elimination of both public and private oversight, we find that their combined effect on contract performance is ambiguous: we observe a decrease in time performance of 1.8 percent and an increase in cost performance of .6 percent. The straightforward implication of our results is that an effective reform should exempt contractors from public oversight and keep the benefits of the private oversight.

This paper contributes to the growing literature on optimal procurement regulation. This strand can be divided into two branches depending on the focus of the analysis: i) papers dealing with ex-ante regulations through the analysis of auction formats, contract types, awarding procedures and their effects on participation and performances (recent examples in this literature include Marion (2007), Board (2007), Marion (2009), Krasnokutskaya and Seim (2011), Bajari and Lewis (2014), and Branzoli and Decarolis (2015)); and ii) papers focussing on ex-post tools for enhancement of contract outcomes, such as oversight (Calvo, Cui and Serpa, 2016) and relational contracting (Banerjee and Duflo, 2000; Calzolari and Spagnolo, 2009). Our paper combines these approaches in disentangling the role of public and private oversight as regulatory elements on the one hand and as a mean to increase monitoring

[13]The premium is incorporated into the bid and affects the probability of winning the tender. Thus, the higher the risk for the surety, the higher the premium charged and the lower the chance of winning.

[14]The definition of *competence* is controversial and we will not address it in the present paper. In our exercise we will proxy competence through the closely related concept of *performance persistence*: we will use a weighted distribution of past contractual performance and divide our sample into competent and incompetent offices depending on the median value (Decarolis et al., 2018).

of contractors on the other. In doing so, we propose direct measurements of outcomes for government purchases. Examples of empirical economic analyses of government efficiency that make use of direct measurements of outcomes include Di Tella and Schargrodsky (2003), Reinikka and Svensson (2004), Olken (2006, 2007), Bertrand, Duflo and Mullainathan (2004), Fisman and Gatti (2006), Fisman and Miguel (2007), and Ferraz and Finan (2008, 2011).

In emphasizing the role for public oversight, we join the discussion of the impact of buyer's discretion in procurement, whose evidence is mixed and hinging on the award procedure employed in the tender. Although increasing buyers' discretion may be beneficial when quality is only partly contractible (e.g., Coviello, Guglielmo and Spagnolo (2017)), rule-based competitive auctions are, in general, believed to be less prone to inappropriate and inefficient practices. One reason for this belief is that in such auctions, there are explicit criteria for selecting the winner. In their paper, Bandiera, Prat and Valletti (2009) identify the amount and the sources of public waste in Italian public procurement. They find that inefficiency is by far the most important dimension in explaining public waste, with heterogeneity across different buyers, and that the best performance - both in terms of active and passive waste - is associated with more discretion. According to Kelman (1990) and Kelman (2005), an ultimate cause of passive waste in the U.S. federal government is that an excessive regulatory burden may make procurement cumbersome and increase average prices: our results on the public oversight effect provide support to this argument. Szucs (2018) exploits the time variation of an Hungarian policy reform to estimate the effects of increased buyer's discretion and find that discretion increases the price of contracts and decreases the productivity of contractors. Hyytinen, Lundberg and Toivanen (2018) provide an analysis of the consequences of a regime change that induced a shift away from discretionary beauty contests to a more rule-based procurement environment, in which only first-price sealed-bid or scoring auctions were allowed. The procurement costs were similar before and after the regime change.

In addition, this work takes part in the debate on the effectiveness of private provision of public goods. In the pursuit of their private interests, the sureties back up the effective completion of public construction projects. By providing empirical evidence for the impact of private oversight on performance outcomes we are thus investigating whether outsourcing a public-interest task to the private sector proves to be beneficial (Banerjee et al., 2017; Hart, Shleifer and Vishny, 1997).

To the best of our knowledge, this paper is the first to empirically assess the role of performance bonding and the associated private oversight of public procurements. Despite not being widely known, performance bonding is a founding pillar of the U.S. public construction procurement, which is a crucial economic sector worth approximately $32 billion, and was extensively used during the recent financial crisis as a fiscal policy tool to stimulate the economy (see the *American Recovery and Reinvestment Act*).[15] Both at the federal level (Miller act) and the state level (Little Miller Acts), there were only slight variations in the regulations before the 2011 reform; therefore, assessing the effectiveness of performance bonds has essentially been impossible. On top of that, the low default rate of federal construction contractors (less than 1 percent) has been interpreted at times as an indication that performance bonds are redundant and represent an unnecessary cost for firms and public buyers, and should therefore be eliminated (Gransberg, Kraft and Park, 2014). This paper, instead, uses novel variation to identify the causal effect of this instrument and reveals that its quantitative effects on contract performance are large and positive, both in terms of time and costs. Furthermore, providing evidence in favor of the screening role of sureties reverses the causality previously highlighted: performance bonding is what helps keep the default rate low by enhancing the selection of the best contractors.

The remainder of the paper unfolds as follows. In section II, we present the concept of performance bonding and the related U.S. legislative context; section III deals with the theoretical background underlying our analysis; section IV outlines the data we employ in our analysis; section V addresses the empirical analysis, outlines the identification strategy and presents results plus robustness checks; in section VI, we discuss the main drivers of our findings. Section VII concludes.

[15] Year 2013, source: Federal Reserve Bank of St. Louis.

II Context

In this section, we first describe the institution of performance bonding and the economic rationale underlying its provision in public procurement regulations; when presenting its legislative foundations in the U.S. federal procurement we define the private oversight. We then shift the focus to the U.S. federal procurement regulation to define and discuss the public oversight.

II.1 Private and Public Oversight

Performance Bonding Procuring supplies entails strategic considerations on competition, tender design and optimal ex-post rating in order to ensure the maximum benefit for the procurer. When dealing with procurement of services, buyers also face uncertainty related to production cost and business factor dynamics: unexpected negative shocks could hit contractors during the execution of the work, leading to profit erosion and, ultimately, losses.[16] In the worst-case scenario, contractors are forced to declare bankruptcy, leaving the work incomplete and the buyer with no party to make claims against. Avoiding such lose-lose outcomes is a first-order concern for all parties, and situations of this sort are typically handled by renegotiating contract provisions either in terms of delivery time or costs. This leaves room for moral hazard and adverse selection issues, as low-quality firms may take advantage of cost uncertainty at the awarding stage, underbid and then renegotiate once awarded the contract - e.g. by pretending to have suffered an unexpected negative cost shock (Guasch, Laffont and Straub, 2008).

Hence, when the contractors' probability of default is high, it makes sense for buyers to take out an insurance to avoid bearing all risks on their own. The performance bonding is a specific line of insurance based on the issuance of a *performance bond* and involving

[16] According to the OECD, *services* are "outputs produced to order and which cannot be traded separately from their production". A broader definition provided by the management literature is based on "the five I's": Intangibility, Inventory, Inseparability, Inconsistency and Involvement. Either way, throughout the paper we will distinguish supplies contracts from service contracts according to the underlying timing of production: while goods could - in principle - be stored and sold outright, services are customized and need time to be produced and delivered after the contract award.

three parties: the surety guarantees that the contractor will perform the tasks demanded by the buyer.[17] In other words, the performance bonding works as a risk-transfer mechanism between the buyer and the surety company, but it is demanded by a third party - the contractor - that guarantees the performance of an obligation.

Prior to issuing a bond, the potential contractor is subject to a *screening process* by the surety - consisting of an assessment of its entire business operations, financial resources, experience, organization, backlog, profitability and management capability - aimed at extrapolating private information on its type. The surety, thanks to its access to firms' information during the prequalification phase and its prior experience of the market, is able to evaluate the contractor's ability to fulfill the contract provisions.[18] The whole process culminates in the determination of a *premium*, an actuarially based fee that varies depending on the size, type and duration of the project and, notably, on how the characteristics of the contractor that emerged from the screening process match the project complexity. In the U.S., the bond price mostly ranges 0.5-3 percent of the contract amount and the potential contractor typically incorporates the bond premium amount into the offer. Hence, the screening enacted by the surety makes the premium a prominent *ex-ante* mechanism for discriminating among potential contractors. The premium, by reducing the asymmetry of information and affecting the offered bid amount, takes a relevant role in determining the quality of the winner in a competitive tender and shifts adverse selection issues away from the procurer, whose only piece of information about the sellers at the award stage is the offer placed.

Furthermore, sureties systematically gather and analyze information regarding bonded contractors *after* the contract award. They have the legal right to access information on work progress, payments and the estimated percentage of completion for bonded projects. Prior to modifying any contractual term, procurers and contractors shall obtain the consent of the surety on the basis of the gathered information on contractor conduct.[19] Hence, in addition to being screened, bonded contractors undergo an *ex-post* monitoring process by sureties. Hence, performance bonding also shifts moral hazard issues away from the buyer.

[17] The legal definitions for buyer and contractor are *obligee* and *principal*, respectively.

[18] Performance bonds are common across the entire U.S. construction industry. Construction bonds generate two-thirds of total surety premia written and 70 percent of total revenues.

[19] Federal Acquisition Regulation (FAR) Part 28.

A comparison with letters of credit (LOC), widely used in the European procurement market, might be useful to better understand how performance bonds differ from other traditional forms of guarantee in their nature and the underlying incentives provided. A LOC, normally issued by a bank, is a cash guarantee to the buyer who can call on demand and receive a pre-specified amount of money if some breach of contract were to occur. A performance bond protects the buyer from nonperformance and financial exposure, should the contractor default. Hence, while the performance of the contract has no or little relation to the bank's obligation to pay on the LOC, the primary focus of a performance bond is the effective accomplishment of the work. The two instruments also differ with respect to their effect on the contractor's borrowing capacity and the prequalification process. In order to issue an LOC, the bank always requires the contractor to pledge specific assets to be paid in case of insolvency. An LOC thus diminishes the contractor's line of credit and appears on financial statements as a contingent liability. The bank examines the quality and liquidity of the asset by checking whether it could back up the debt; if this is the case, no further prequalification is required. Hence, a bank issuing an LOC takes no risk and has no incentive to screen the contractor, whose liquidity is reduced to back up the LOC. Should the applicant be unable to make payment on the purchase, it shall cover the outstanding amount. In contrast, performance bonds are issued on an unsecured basis and neither alter firms' assets nor diminish the contractor's borrowing capacity; in other words, the surety bears part of the project risk. In order to ensure the delivery of the contract object in case of contractor's default the surety has to choose between the following: (i) covering production costs by itself and allowing the contractor to finish the works; (ii) selecting a new contractor to conclude the residual tasks; or, only as a last resort, (iii) refunding the bond value to the buyer, leaving the execution incomplete.

These crucial differences imply that an LOC is likely to be unavailable to companies with few assets, which excludes them from participating in the tender and thus reduces competition on dimensions not related to quality. Since sureties, which must have sufficient assets to back up the bonds they issue, are partially responsible for the completion of the works, they have strong incentives to properly screen potential contractors and to assess their ability to execute the job. This point crucially inspired our work. *Ceteris paribus*, a bonded project is more likely to be completed in accordance with the contract provisions as the likelihood of contractor default or any breach of procurement contract clauses is reduced, while the awarding price may be higher due to a premium.

Performance bonds are required for US Government procurement by the Miller Act.[20] The Act applies *only* to contracts awarded for the construction, alteration, or repair of any public building (for the sake of simplicity, we will refer to this subset of contracts as *constructions* henceforth) of the U.S. federal government. The Miller Act imposes that, in order to be allowed to participate in the tender, potential contractors must furnish the federal government with a performance bond pre-approval. Typically, the performance bond amounts to the 100 percent of the contract price.[21] Throughout this paper we refer to the performance bonding as *private oversight*.

Public Oversight The Federal Acquisition Regulation (FAR), the guidebook governing the public procurement process in the U.S., provides a set of rules that contracting offices are to comply with during both the awarding phase and the operation phase of the acquisition. Following Calvo, Cui and Serpa (2016), and as opposed to the above presented private counterpart, we refer to these formal background rules collectively as *public oversight*.

Contracting officers are required by the FAR to use one of the two following formal solicitation methods when acquiring supplies or services: sealed bidding or negotiation. They involve six and nine formal steps, respectively, each requiring a series of checks on bidders' documentation enacted by the contracting officer before awarding the contract - i.e., ex-ante.[22] Moreover, during the operational stage - i.e., ex-post - contracting officers require sellers to complete expenditure justification forms and submit cost/pricing data (in order to certify that expenses are based on adequate price competition); eventually, sellers must

[20] 40 U.S.C. sections 3131-3134

[21] Contractors are free to choose their own surety from a list of financial companies which the U.S. Department of Treasury establishes as qualified to underwrite performance bonds on federal government projects. This certificate of authority also determines the amount of the maximum limits of coverage for each of these. In other words, a surety that wants to issue bonds for federal government construction projects is in turn subject to a financial review that officially sets its bond size limit.

[22] Refer to the appendix, section A.3, for a detailed review of the procedures.

submit reports on project progresses to specific evaluation teams.[23]

II.2 Simplified Acquisition Procedures

Federal Acquisition Streamlining Act The Simplified Acquisition Procedures, introduced with the Federal Acquisition Streamlining Act of 1994, aim at reducing the administrative burden for the sellers, mainly small businesses, when working for the Government.[24] Under the simplified acquisitions, federal buyers need not allocate time and resources to the formal acquisition procedures described above - i.e., they do not have to exert any public oversight.[25] Contracting officers are encouraged to use the simplified acquisitions, and thus to exempt contractors from public oversight, to the maximum extent practicable for purchases of supplies or services whose anticipated dollar value does not exceed a monetary cutoff - the Simplified Acquisition Threshold.[26] The private oversight (i.e., the performance bonding) applies above the same cutoff and, for the sake of convenience, we refer to a single *oversight threshold* for the implementation of both regulations. More specifically, a contractor awarded a construction (non-construction) project whose anticipated value lies above the oversight threshold is subject to both public and private oversight (public oversight only).

[23]The number and type of checks are similar for each contracting office, as provided by the FAR, and their scope is analogous; we can coherently group them into one set.

[24]FAR part 13.

[25]There are five buying methods prescribed in FAR Part 13 for simplified acquisition purchases. The two major methods are Purchase Orders and Blunket Purchase Agreements. A Purchase Order (FAR Part 13.302) is a commercial document issued by a buyer to a seller, indicating types, quantities, and agreed prices for products or services the seller will provide to the buyer. Sending a PO to a supplier constitutes a legal offer to buy products or services. Acceptance of a PO by a seller usually forms a one-off contract between the buyer and seller, so no contract exists until the purchase order is accepted. A Blanket Purchase Agreement (BPA) is a simplified method of filling anticipated repetitive needs for supplies or services by establishing "charge accounts" with qualified contractors. BPAs should be established for use by an organization responsible for providing supplies for its own operations or for other offices, installations, projects, or functions. The use of BPAs does not exempt an agency from the responsibility for keeping obligations and expenditures within available funds and executed in accordance with Federal Acquisition Regulation (FAR) 8.405-3.

[26]Indeed, according to the FAR, the purpose of the Simplified Acquisition Procedures is to reduce administrative costs, improve opportunities for small and disadvantaged businesses to obtain a fair proportion of government contracts, promote efficiency and economy in contracting; and avoid unnecessary burdens for agencies and contractors.

Exogenous Variation We exploit a change in the oversight threshold that was enacted in October, 2010, to inform our identification strategy. The 41 USC 1908 requires the government to review the acquisition-related thresholds every five years - for inflation. Notably, "to review" does not necessarily imply "to adjust": the choice to move thresholds depends on several factors other than the change of the Consumer Price Index in the previous five years, including political and economic considerations. The law applied to both the simplified acquisitions and the Miller Act provisions through an update of the oversight threshold, which was raised from \$100,000 to \$150,000 on October 1, 2010.[27] The same thresholds, although reviewed in accordance with the law provisions, were not changed in 2005 and 2015.

Figure (1) provides a stylized timeline of the outlined framework. Left panel represents the fiscal years 2005-2010[28] for construction and all other contracts ("non-construction" from now on), right panel refers to the period 2011-2015, for the same contracts. The horizontal dotted line represents the oversight threshold, moving upward as of FY 2011; the grid identifies awarded contracts worth \$100,000 to \$149,999, while the background colors refer to oversight application (gray, dark and pale) or exemption (white). In the case of construction contracts, the exemption included both public and private oversight, while for non-construction the exemption was from public oversight only. Over the time span considered, construction contracts valued above \$150,000 (below \$100,000) are always (never) subject to private and public oversight, while non-construction contracts of the same amount are subject to (exempted from) public oversight only.

III Theoretical Background

While helping to scrutinize among potential contractors and to restrain vendors' misconduct, public oversight introduces a burden in terms of both time and cost due its intrinsic characteristics. In order to comply with solicitation rules and produce the required paperwork, sellers must divert resources away from contract-specific tasks, and their leeway is hampered

[27] The adjustment is rounded - in the case of a dollar threshold that is not less than \$100,000, but is less than \$1,000,000 - to the nearest \$50,000.

[28] Fiscal year 2010 ends on October, 2010. The threshold revision was enforced from fiscal year 2011 on.

Table 1: Reform Timing

	Pre-2011		Post-2011	
	Construction	Non-Construction	Construction	Non-Construction
Above $150k	Private and Public Oversight	Public Oversight	Private and Public Oversight	Public Oversight
$100-150k	Private and Public Oversight	Public Oversight	None	None
Below $100k	None	None	None	None

Notes: Contracts subject to (gray) and exempted from oversight before and after October, 2010. The $100,000-150,000 class (grid) identifies the treatment group, i.e., those contracts subject to oversight before but not after the reform. Upper control group - "Above $150k" - includes contracts always exposed to oversight (i.e., always gray, both construction and non-construction) while the lower control group - "Below $100k" - consists of contracts never exposed (always white).

by the need for public approval. This is extensively recognized by FAR itself when introducing the simplified acquisition procedures.[29] To sum up, enforcing public oversight may lead to two conflicting phenomena:

- Hypothesis *a.1)* : The introduction of an unnecessary bureaucratic burden for contracting parties causes longer delays and higher costs - *red tape effect*;[30]

- Hypothesis *a.2)* : formal solicitations procedures reduce discretion of contracting officers at the award phase and project supervision reduces the risk and the extent of opportunism, slack conduct or misbehavior in contract execution - *public adverse selection and moral hazard effect*.[31]

An ex-ante assessment of the effect of public oversight on contractors' performance is not

[29] The purpose of FAR part 13 is to prescribe simplified acquisition procedures in order to i) reduce administrative costs; ii) improve opportunities for small, small disadvantaged, women-owned, veteran-owned, HUBZone, and service-disabled veteran-owned small business concerns to obtain a fair proportion of Government contracts; iii) promote efficiency and economy in contracting; and iv) avoid unnecessary burdens for agencies and contractors.

[30] See Bozeman (1993) for a review of the theory of red tape and public contracting.

[31] See Spiller (2008) for the theory on public contracts and opportunism; see also Decarolis, Pacini and Spagnolo (2016).

trivial. On the one hand, the two effects are competing;[32] on the other hand, both might be non-linear in the contract amount - monitoring might be a wasteful activity only for small projects, and could lead to savings for larger ones.

On top of public oversight, firms competing for federal construction contracts are *also* required to obtain performance bonds and be subject to private oversight. This entails oversight exerted by private companies, i.e the sureties. In turn, the effect of performance bonding on contract outcomes may have two sources:

- Hypothesis *b.1)*: firms subject/not subject to Miller Act provisions are structurally different due to the screening effect induced by sureties - *private adverse selection effect*;
- Hypothesis *b.2)*: as for public oversight, being covered and monitored by a surety gives firms more incentives to complete contracts under the terms and conditions agreed - *private moral hazard effect*.

Hypothesis *b.1)* is the one proposed by Calveras, Ganuza and Hauk (2004) (CGH henceforth), according to which we should observe a different pool of winning firms before and after the reform.[33] Specifically, since sellers are no longer subject to the pre-bidding screening process, we should observe a high turnover rate between firm types. After the reform, low-quality firms are supposed to be more likely to win at the expense of the good types given that their low quality does not reflect on higher premia charged by sureties anymore.[34] Thus we would expect more bad-type contractors to enter the pool of winners, good types to exit and the quality of the average contract outcome to decline accordingly. Hypothesis *b.2)* underlies a different prediction on the pool of winning firms. The assumption that surety companies do not screen potential contractors through a premium discrimination implies that we should not observe any significant change in the composition and structure of

[32] Identifying the extent to which the red tape and the moral hazard effects induced by the public oversight interact and affect the contract outcomes goes beyond the scope of the present paper.

[33] The FPDS-NG only reports tender winners. Indeed, according to hypothesis *b.1)*, the pool of potential contractors does not necessarily change with or without screening.

[34] In CGH terms limited liability companies are more willing to bid aggresively and, ultimately, face risks and an unexpected need to revise contract terms.

awarded firms after the reform. In such a framework, what matters instead is that removing performance bonds reduces the incentives for the *same* firms to exert the effort required to accomplish the contract tasks. To guarantee contract completion, sureties check the status of works and evaluate contractors' performance. In their absence, an issue of moral hazard arises and contractors tend to perform worse.

Hypotheses *b.1)* and *b.2)* are not competing and we expect both to be relevant in the public procurement market. The role designed by the law for surety companies is meant to minimize both effects through an ex-ante and ex-post monitoring of contractors. The overall effect of private oversight on contract outcomes amounts to the sum of selection, monitoring and the interactions of the two, and we expect it to be positive in terms of contract outcomes.

IV Data

IV.1 FPDS Dataset

The data we use are sourced from the Federal Procurement Data System (FPDS), a database to which federal contracting officers in the U.S. must submit complete reports on procurement contract actions, as required by the FAR. It contains all contracts, both supply- and service-based, that have been awarded by the U.S. government and exceed an individual transaction value of $2,500, as well as every following activity.[35] The dataset also includes several variables related to the transaction itself, including buyer and seller characteristics in addition to solicitation and contract information, such as the signature, award and insertion dates, the contract object and its category (i.e., service or supply).

Importantly, we observe the type of solicitation procedures used, which reveals whether a contract is awarded through Simplified Acquisition Procedures (i.e., no oversight) or other procedures (sealed bidding or negotiation). Using this information, we build the binary variable SAP_i, indicating whether contract i has been waived from public and private oversight

[35] Data are gathered by contracting offices in 23 agencies. In Tables (A.3) and (A.4) we report the number of contracts per agency/year.

or not. The SAP variable crucially supports our identification strategy: ideally, we would like to observe the engineers' estimated value (EV_i), which is the piece of information used by the contracting office to assign the public oversight treatment to a contract. However, this is not recorded into FPDS and we are able to overcome the issue only combining information provided by i) SAP, that is we identify contracts exempted from public and private oversight, and ii) the ex-post contract value.[36] The version of FPDS employed dates back to September, 30 2015.

IV.2 Data Management

We split the data into two main groups: contracts and amendment records. The former refer to the first transaction between a procurer and a vendor and correspond to our unit of observation, whose reported characteristics represent the benchmark procurement agreement information. The latter account for all the revisions, modifications or corrections to existing contracts. Each contract is identified through a unique ID which is used to mark all its present and future alterations; therefore, we are able to track the entire contract history and link each contract to its revisions. Amendment records are classified according to the reason for contract modification, which is reported alongside the extra cost and time taken to complete the works. We further group them into *in-scope* or *out-of-scope* revisions, depending on whether the goal of the amendment is consistent with the initial contract terms.[37] We use the *in-scope* amendments to build the outcome measures of our empirical analysis presented below.[38]

[36] Consider two contracts, A and B, whose observed contract value is $105,000, both awarded before the threshold revision. However, the unobservable engineers' estimated value of A, EV_A, is $110,000, while $EV_B = \$95,000$. According to the contract value, they are both subject to oversight. However, exploiting the fact that $SAP_A = 0$ and $SAP_B = 1$, we can proceed to the correct identification and avoid any source of bias in the estimates.

[37] According to the FPDS data dictionary, we label as *out-of-scope* all amendments classified as "Additional Work (new agreement, FAR part 6 applies)", "Novation Agreement", "Vendor DUNS or name change - Non-Novation" and "Vendor Address Change". We consider all other amendments as being within the scope of the project.

[38] Before initiating a modification, the contracting officer must determine if the proposed effort is within the scope of the existing contract or is a new acquisition outside of the scope. A new requirement outside of the scope of the existing contract must be processed as a new acquisition. Contract scope means, in simple terms, that the contemplated change must be generally related to the work originally contracted for. If a contract was awarded for the design (and only the design) of an automated information system, it could not

Performance Indexes First, we define: i) *Time Overrun*, representing the days in excess of a project's initial deadline; measured as the difference between the actual completion date and the estimated one and ii) *Cost Overrun*, standing for the expenses in excess of a project's initial budget; it is the sum - in thousands of dollars - of all renegotiated amounts. *Time Overrun* and *Cost Overrun* are widely used proxies for contractual performance;[39] however, there are circumstances in which renegotiating the contract terms leads to optimal outcomes - typically, this is the case for complex, structured projects likely to be subject to unexpected events (negative cost shocks, adverse natural conditions, etc.). Given high-value contracts are the minority in our sample, and according to Spiller (2008), who argues that renegotiations are suboptimal in the public procurement context, we consider the measures built on *in-scope* amendments only to adequately reflect the performance of a contractor.[40]

In order to compare the two overrun measures with the initial expected outcomes - that is, the time/cost of completion specified in the contract terms - we specify two indexes for contract performance like:[41]

$$performance_{ig} = \frac{expected\ outcome_{ig}}{expected\ outcome_{ig}\ +\ overrun_{ig}}$$

where i refers to the contract and $g = [time, cost]$. By construction, it maps the couple $[expected\ outcome_{ig}\,; overrun_{ig}]$ to the interval $[0, 1]$, with an increasing performance approaching 1, that is in the case of no overruns. Not surprisingly, the two performance measures are positively correlated (50 percent).[42]

be later modified to have the contractor provide and install hardware.

[39] Among the others, see Lewis and Bajari (2017), Coviello, Guglielmo and Spagnolo (2017), Decarolis (2014) and Guasch, Laffont and Straub (2008).

[40] Spiller (2008)'s argument unfolds as follows: given the formal, bureaucratic nature of public contracting, any terms renegotiation would add adjustment costs, providing weaker incentives to adapt for both contractors and public authorities. Bajari, Houghton and Tadelis (2014) provide support to this hypothesis by quantifying in 8 to 14 percent of the winning bid the adaptation costs in their construction data.

[41] The two performance measures are positively correlated (48 percent). This feature of our data differs from that in Decarolis (2014), who finds a nearly zero correlation between time and cost renegotiations and no evidence of a nonlinear relationship. He stresses, however, that designing the contract in such a way that the contractor would be in charge of both the design and the execution of the project would lead to shorter time and greater cost overruns. We are not able to reproduce his results since the FPDS does not contain such information.

[42] Figure (A.9) in the appendix is a scatterplot showing the correlation between *cost performance* and *time performance*.

Also, we build two binary variables indicating whether the contract terms have been amended, that is at least one modification follows the initial contract signature in terms of completion time (*Time Amended*) or final cost (*Cost Amended*).

We also store the average amount of time (*Average Time Overrun*) and cost (*Average Cost Overrun*) overruns - i.e., $\frac{\sum_{k=1}^{K} amount\ amended_{i,k}}{number\ amendments_i}$, where i stands for the contract and k the amendment. These variables are defined only for the subset of contracts subject to at least one revision.

The FPDS dataset includes a number of other variables from which we build the controls in our regressions. *SAP* is a binary variable indicating whether the contract has been subject to the Simplified Acquisition Procedures; *Constr* is an indicator for construction contracts; *Fixed Price* indicates whether contracts are priced with a fixed price or cost plus format - i.e., if the supplier is paid a fixed amount, regardless of costs incurred, or if is entitled to obtain compensation in proportion to its costs plus a mark-up; *Small* is an indicator for small business vendor;[43]; *Negotiation* is a dummy variable for contracts awarded through negotiated procedures; and *Bureau Size*, which is the cumulative value of contracts a bureau has awarded in the current year for the same service or construction category.

IV.3 Sample Selection

We restrict our sample to those contracts awarded through competitive solicitations because the effect of the treatments would otherwise not be observable.[44] For similar reasons, we focus on contracts whose tasks are such that the vendor can influence the outcome metrics through effort. Supply contracts do not allow for renegotiations. Hence, for these contracts our measure of performance does not proxy outcome quality whatsoever and we exclude them from the analysis.[45] The same rationale applies to the service subcategory "Lease or

[43]The Small Business Authority (SBA) labels small firms based on the particular service category which the contract belongs to, and on sellers' characteristics (revenues, number of employees, etc).

[44]We consider as competitive a lot for which the extent of competition is labelled "Full and open" and whose participation is not set aside to any specific group of firms. In non-competitive tenders, the participation criteria restrict the competition *ex-ante* to dimensions other than quality (e.g. Athey, Levin and Seira (2011))

[45]The typical supply contract shows a 0 value in *time/cost overruns* and a unit value in both performances.

Rental of Equipment, Structures, or Facilities".[46] In order to keep a balanced time-window around the SAT update, we rule out observations before January, 1, 2005, and cover the years 2005 to 2015. We eliminate contracts whose expected termination date is beyond the date of data download - September 30, 2015 - to keep only completed projects. We also drop contracts related to certain commercial items that make use of simplified procedures for the acquisition of services for amounts greater than the oversight threshold. This cleaning process yields a sample of 226,161 contracts and 23,870 unique firms .[47]

Two sets of contracts - the two solid colored sections in Figure (1) - are potential candidates for use as control groups: the "always exposed" set (*upper control group*) and the "never exposed" set (*lower control group*). The reform date and the two treatments cluster the sample into 6 distinct groups: the treatment group, counting all contracts - constructions included - valued between \$100,000 and \$149,999 that are subject to public oversight before but not after the reform; upper and lower control groups, consisting of all contracts valued more than \$150,000 or less than \$100,000, respectively; and construction treatment, upper and lower control subgroups, including construction contracts only, subject to private as well public oversight, with the same monetary cutoffs.

In Table (2) we report summary statistics for the Service treatment group and upper control group, both before and after the reform.

V Empirical Analysis

In this section, we first explain the econometric strategy used to identify the effect of private oversight and public oversight. Then, we present the estimation results and the relative robustness checks.

[46]Services included in the sample are: Special Studies/Analysis, Not R&D; Architect and Engineering Services; Information Technology and Telecommunications; Purchase of Structures/Facilities; Natural Resources Management; Social; Quality Control, Testing, and Inspection; Maintenance, Repair, and Rebuilding of Equipment; Modification of Equipment; Technical Representative; Operation of Structures/Facilities; Installation of Equipment; Salvage; Medical; Support (Professional/Administrative/Management); Utilities and Housekeeping; Photo/Map/Print/Publication; Education/Training; Transportation/Travel/Relocation.

[47]The firm ID variable is missing in approximately 57 percent of the contracts in our sample.

Table 2: Summary statistics - Service sample

	Upper Control Group							
	Before				After			
	Mean	SD	Median	N	Mean	SD	Median	N
Time Performance	0.7	0.3	0.74	51,245	0.7	0.3	0.82	52,737
Num Time Amendments	2.4	2.8	1	51,245	1.4	2.1	0	52,737
Prob Time Revision	0.6	0.5	1	51,245	0.6	0.5	1	52,737
Avg Time Overrun	199.4	252.3	128.6	27,887	214.0	287.9	115.6	22,684
Cost Performance	0.7	0.3	0.91	51,245	0.8	0.3	0.98	52,737
Num Cost Amendments	2.8	3.2	2	51,245	1.8	2.6	1	52,737
Prob Cost Revision	0.6	0.5	1	51,245	0.5	0.5	1	52,737
Avg Cost Overrun	273.8	333.5	205.1	31,587	307.8	413.7	176.8	27,481
Contract Value	1,896.8	9,311.5	421.7	51,245	1,098.6	4,334.6	300	52,737
# Contractual Days	417.6	404.5	364	51,245	305.1	222.1	357	52,737
Offers received	5.4	18.7	2	51,245	5.7	18.7	2	52,737

	Treatment Group							
	Before				After			
	Mean	SD	Median	N	Mean	SD	Median	N
Time Performance	0.8	0.3	1	13,774	0.8	0.3	1	2,061
Num Time Amendments	1.1	2.0	0	13,774	0.7	1.5	0	2,061
Prob Time Revision	0.4	0.5	0	13,774	0.3	0.5	0	2,061
Avg Time Overrun	229.9	287.3	141	4,763	187.6	250.5	100	565
Cost Performance	0.8	0.3	1	13,774	0.9	0.2	1	2,061
Num Cost Amendments	1.3	2.3	0	13,774	0.8	1.7	0	2,061
Prob Cost Revision	0.4	0.5	0	13,774	0.3	0.5	0	2,061
Avg Cost Overrun	101.2	125.9	64.2	5,377	67.5	96.7	34.1	625
Contract Value	122.2	15.2	121.2	13,774	122.0	16.1	120.1	2,061
# Contractual Days	298.9	342.8	215	13,774	228.2	191.1	205	2,061
Offers received	4.7	14.5	2	13,774	3.9	18.9	1	2,061

Notes: the table reports descriptive statistics for both the upper control group (above panel) and the treatment group (below panel), before (left side) and after (right side) the threshold revision. *Time* and *Cost Performance* are relative measures of performance - bounded 0 to 1; *Num Time* and *Num Cost Amendments* count the number of amendments per contract, while the relative *Prob* is a binary variable that takes value 1 in case of any amendment occurs; *Avg Time* and *Avg Cost Overrun* account for the average extra time or extra cost and is defined only for contracts which had at least one amendment; *Contract Value* is expressed in US$ thousands; *Offers Received* report the number of offers received per tender.

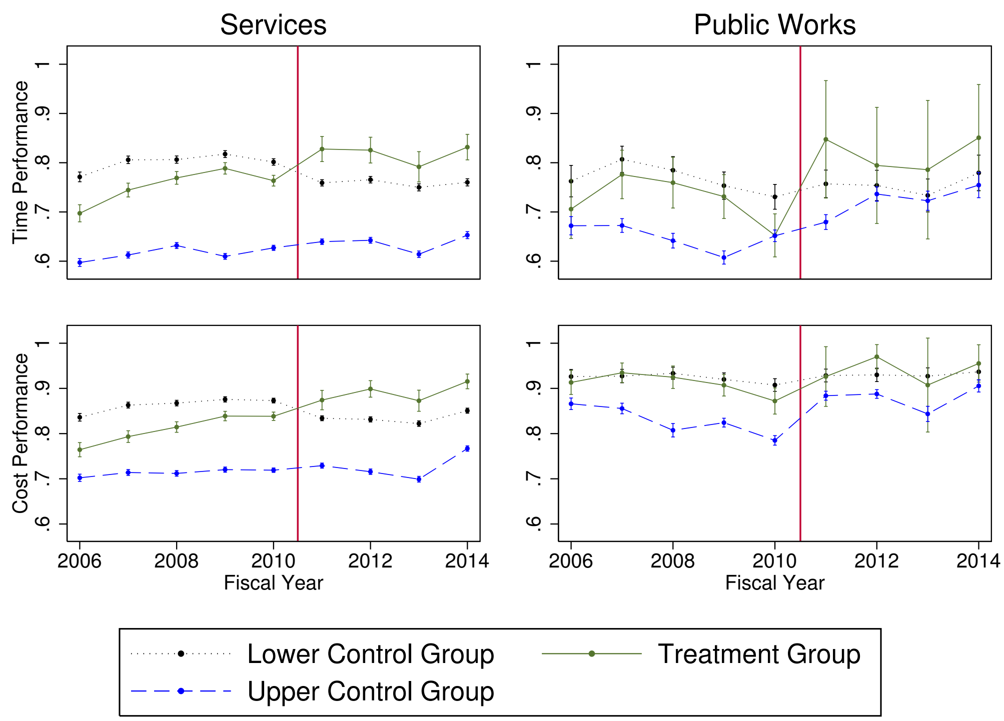
V.1 Identification strategy

We shall exploit the threshold adjustment in order to separately identify the effect on performances of public oversight and private oversight exemption. In principle, we would want to randomly assign the provisions across solicitations and perform a pairwise comparison of the average outcomes of the groups in the two cases. In the absence of a controlled randomized trial, we are forced to turn to non-experimental methods that mimic it under reasonable conditions.

Construction contracts above the oversight threshold are exposed to both public and private oversight, while non-construction contracts are subject only to public oversight. Hence, the grid in Figure (1) identifies the treatment group: for construction contracts, the treatment results in exemption from both types of oversight, while for non-construction it is the exemption from public oversight only. Sections without grids identify upper (gray) and lower (white) control groups. We start by considering all contracts and present a plain difference-in-differences (DD) strategy. We then focus on the construction/non-construction distinction and discuss how to nest two DD analyses through the difference-in-difference-in-differences (DDD, or triple difference) approach.

The simplest framework for a DD estimation requires a set of individuals observed over two periods. A subset of observations - the treatment group - is exposed to a treatment in the second period; the other subset - the control group - is never exposed. Measuring the difference in the average outcome between the groups, while keeping everything else constant, yields the average treatment effect on the treated. The underlying assumption, which crucially informs the DD identification, states that the difference in expected outcome between the groups is constant across periods, conditional on observables; in other words, one assumes the trends of the variable of interest in the two groups would have been parallel had the treatment not occurred. In our setting, the contract is the unit of observation, the treatment is the waiver of oversight, the periods are determined according to the reform date, and the groups are defined as above. To verify whether the parallel trend assumption is reasonable in our data, we plot the yearly average time series of time performance and cost performance for the Service treatment and both control groups in Figure (1).[48] The trends appear to be parallel throughout the pre-treatment period.

Figure 1: Time and Cost Performance: Yearly Averages



Notes: Trends in yearly averages of *Time Performance* (above panels) and *Cost Performance* (below panels) for treatment, upper and lower control groups. Left panels refer to the service contract sample, right panels to the constructions sample. The vertical line corresponds to October, 2010.

If the parallel trends assumption holds, it is then possible to identify the average treatment effect on the treated by running a linear regression:

$$Y_{it} = \beta_1 D1_{it} + \beta_2 D2_{it} + \theta_a (D1_{it} * D2_{it}) + \epsilon_{it} \quad (1)$$

where $D1_{it}$ and $D2_{it}$ are binary indicators for group (treatment/control) and period (before/after), respectively. The term $(D1_{it} * D2_{it})$ identifies the treatment and its parameter

[48] Cost performance shows a sharp increase in both the treatment and the control groups. This is possibly due to the presence of incomplete contracts in our sample when approaching the date of download. We show that the results are robust to the narrowing of the estimation time window.

$\hat{\theta}_a$ amounts to the average treatment effect on the treated. In our setting, in which one treatment is nested onto the other, however, $\hat{\theta}_a$ is biased and the very definition of treatment is ambiguous, as it encompasses effect of the waiver of both public and private oversight. The latter is relevant to treated construction contracts only, but its effect is estimated jointly on the whole sample and cannot be disentangled via a plain DD.

In order to deal with two nested treatments, we rely on an augmented version of the DD. The triple differences approach nests two DD models like (1) in a single equation and, controlling for the relative differences between treatment and control groups, consistently estimates the average treatment effects.[49] Specifically, starting from equation (1) we define $D3$ as an indicator variable for the subset of individuals subject to the second treatment and augment the model with another tier of differences:

$$\begin{aligned} Y_{it} = \alpha + \beta_1 D1_{it} + \beta_2 D2_{it} + \theta_a(D1_{it} * D2_{it}) + \beta_3 D3_{it} + \\ + \beta_4(D1_{it} * D3_{it}) + \beta_5(D2_{it} * D3_{it}) + \theta_b(D1_{it} * D2_{it} * D3_{it}) + \epsilon_{it} \end{aligned} \tag{2}$$

In equation (2) the triple interaction term $(D1_{it} * D2_{it} * D3_{it})$ marks the individuals subject to both treatments. In our framework, the coefficients of interest $\theta_i, i \in [a, b]$ capture the effect of the waiver of both types of oversight. As in the case of the plain DD, these are identified as the difference between the observed effects of treatment on the treated and the counter-factual outcome in the absence of treatment, which is assumed to be parallel to that of the control group.

Intensive margin We treat our data as a pooled cross-section and use *upper control group* in the baseline and main robustness specifications.[50] In the core analysis of the paper, we examine the treatment effects on the intensive margin; more specifically, we estimate a DDD on cost and time performance metrics. Indicating the contract outcome variable by Y_{ijt}, we

[49] See Berck and Villas-Boas (2016), among others, for further details.

[50] The population of construction contracts always subject to both public and private oversight.

specify the following linear equation:

$$
\begin{aligned}
Y_{ijt} = \alpha + \beta_1 Waiver_{it} + \beta_2 Post_{it} + \theta_{public}\left(Waiver_{it} * Post_{it}\right) + \\
+ \beta_3 Constr_{it} + \beta_4 \left(Constr_{it} * Waiver_{it}\right) + \beta_5 \left(Constr_{it} * Post_{it}\right) + \\
+ \theta_{private}\left(Waiver_{it} * Post_{it} * Constr_{it}\right) + \gamma X_{it} + \zeta_j + \delta_t + \varepsilon_{ijt}
\end{aligned}
\tag{3}
$$

where i refers to the contract, j is the contracting office and t indicates the year. $Waiver_{it}$ is the binary variable marking whether the contract value lies within the treatment band, i.e. oversight FY2005-2010, no oversight FY2011-2015 - and captures differences between the treatment and control groups prior to the policy change.[51] $Post_{it}$ is a dummy variable for contracts awarded after the reform and captures aggregate factors that would cause changes in Y_{ijt} even in the absence of a policy change and the interaction term $Waiver_{it} * Post_{it}$ captures the effect of exempting contracts from public oversight. $Constr_i$ is a binary indicator for construction works and the triple interaction term $Waiver_{it} * Post_{it} * Constr_{it}$ indicates the construction contracts subject to private oversight.[52] Finally, X_{it} are contract- and contractor-specific characteristics at the time of the award and ζ_j and δ_t are contracting office and year fixed effects, respectively. The coefficients of interest are θ_{public}, representing the average treatment effect of the exemption from public oversight, and $\theta_{private}$, capturing the effect of the exemption from private oversight.

In order to fully characterize the treatment effects on the treated, we will analyze both the *intensive margin* - the total and average amount - and the *extensive margin* - the probability - of contract amendment. This approach is crucial to unveil the channels through which contractual performance is affected by the reform.

Extensive margin The triple difference analysis identifies the treatment effects on the intensive margin of outcome measures. In order to fully describe the causal effects of the

[51] Specifically, $Waiver_{it}$ is the interaction between the binary variable indicating whether the contract value lies between 100,000 and 150,000$ and SAP as defined in section IV.2

[52] In terms of equation (2), $Waiver_{it}$ corresponds to $D1_{it}$, $Post_{it}$ to $D2_{it}$ and $Constr_{it}$ to $D3$, respectively.

treatments on the performance, we need to investigate whether treated firms are more likely to renegotiate. More specifically, we are interested in assessing the treatment effects on the probability of amending the contract terms. When not being monitored, firms have more discretionary power during job planning and execution. On the other hand, this leaves room for opportunistic incentives in contract revisions and they may find it more convenient to bargain with the public administration more often at lower amounts. We expect this effect to be even stronger in the construction industry, where the decision to renegotiate with the sponsor must be arranged with the surety, and represents a last resort for contractors. Any minor issue in terms of costs or time could be managed by the surety itself. Hence, in the absence of private oversight, contract revisions become a viable option to overcome unexpected shocks.

We will test these conjectures running a DDD regression of *Time Amended* and *Cost Amended* on treatments and controls. The above premises underlie a second set of conjectures regarding the intensive margin of amendments. If sureties handle minor issues and help contractors to overcome them without contract revisions, we would expect the average overrun to be higher in their presence, since otherwise the sponsor itself has to take care of minor issues. Hence, we proceed with a DDD analysis of the average overrun - *Average Time Overrun* and *Average Cost Overrun* - only for those contracts subject to at least one amendment.

Identification issues and data features The chief concern in our empirical framework is that, as already mentioned in section IV.1, we do not explicitly observe the engineers' estimated value (EV_i). Since we rely on a combination of i) *SAP*, identifying contracts without public oversight, and ii) their award value, we cannot identify the lower bound of treated contracts: this exposes our treatment group sample to the risk of spurious contamination. When testing for robustness of our results, we show that the contract award amount is a good proxy of the engineers' estimate and that the misclassification of contracts to the treatment group is residual.

A very nice feature of our data is that we can run the model on two equally valid sets of control groups: switching from one to the other, as long as the parallel trends assumption holds, should not alter the DDD estimates. In fact, as shown in section V.3, our results are robust to the choice of either group. Finally, it is crucial to remark that contractors decide whether to participate in the tenders and the choice to be subject to the treatment is endogenous. On top of that, the surety company exerts an ex-ante selection on potential contractors, affecting the pool of winners on the quality dimension (see section VI for further details). For all these reasons, a regression discontinuity design (RDD) approach is not a viable option. In order to test for endogenous sorting or discontinuities in the forcing variable, we performed the McCrary (2008) density test for post-law data (see Figure (A.3)). The sharp discontinuity of the running variable at the threshold, highlighted by the graph and confirmed by the highly significant test results, rules out any possibility of running a usual RDD with our data; see Appendix for further details.

V.2 Results

Triple difference Table (3) reports the DDD regression of contract outcomes - time performance in panel (a) and cost performance in panel (b) - on the treatment variables as defined in equation (3). Column 1 reports results of a triple difference model based on equation (3) with no further controls. Specifications 2 to 5 saturate the model by iteratively including controls plus an increasing number of fixed effects (bureau, year, state, and contract category). To deal with a collection of minor problems about normality, heteroscedasticity or observations that exhibit large residuals, leverage or influence, standard errors are estimated using the Eicker-Huber-White estimator.[53]

The public oversight treatment ($\hat{\theta}_{public}$) is robust to the choice of controls and fixed effects. However, adding bureau fixed effects - column 3 - seems to have significant effects on the magnitude of estimates. Similarly, the effect of private oversight waiver ($\hat{\theta}_{private}$) is boosted and becomes statistically significant once accounting for bureau specific features. This is not

[53]Standard errors estimates are robust to various choices of clusterization level. In Table (A.8) we report the estimated parameters of the baseline model for both *Time Performance* and *Cost Performance* with standard errors clustered at different levels.

surprising: as shown in Section VI, persistency of the bureau's performance matters in terms of contract outcomes. On the other hand, controlling for year (column 4), state (column 5) or object fixed effects does not alter results substantially.

Our baseline estimates, in column 5, show that waiving public oversight positively affects contract performance, but the absence of private oversight offsets such gains. Specifically, we find that the waiver of public oversight effect is positive both in terms of time performance (+7.2 percent) and cost performance (+5.3 percent). On the other hand, removing private oversight worsens both measures of performance: it leads to a 9 percent decrease in terms of time performance and to a 4.2 percent decrease in terms of cost performance. The composite effect is ambiguous and depends on the dimension considered. The upward shift of the oversight threshold produces an overall decrease in time performance (-1.8 percent) and a slight increase in cost performance (+0.6 percent).[54]

Extensive and Intensive Margin In Table (4) we present the estimated treatment effect on the extensive margin for time/cost outcomes and confront them with the intensive margin of overrun - i.e., the treatment effect on the average overrun. Odd columns report the results of a linear probability model on an indicator function for amendment while even columns report the estimates of a DDD - equation (2) - on the average overrun. The model employed for each regression includes all fixed effects and controls of column (5) in Table (3). The probability of a time amendment falls in the treatment group (-5.2 percent), while no significant effect is found for the construction treatment subgroup. Results are similar for cost margin, with an estimated decline in the treatment group (-3.1 percent) and no effect in the construction treatment subgroup. The intensive margin analysis yields a similar picture: public oversight causes lower average overruns in terms of both time and cost (-57.5 days and -$42,187, respectively), while private oversight leads to higher average time overruns (+73.5 days) and lower cost overruns (-$67,807). The latter is in contrast with our baseline finding on the private oversight impact on cost performance, although the effect is possibly counterbalanced by the non-significance of the effect at the extensive margin. In words,

[54] We calculate the percent change for constructions only, since both types of oversight apply to these contracts.

Table 3: Triple Difference - Contractual Performances

Panel (a): Time Performance

	(1)	(2)	(3)	(4)	(5)
$\hat{\theta}_{public}$	0.026	0.028	0.050	0.050	0.048
	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)
$\hat{\theta}_{private}$	-0.001	0.004	-0.064	-0.066	-0.067
	(0.025)	(0.025)	(0.024)	(0.026)	(0.027)
N	98,089	98,089	98,089	98,089	98,089
R^2	0.015	0.028	0.094	0.098	0.110
$Avg_{services}$	0.763	0.763	0.763	0.763	0.763
Avg_{works}	0.763	0.763	0.763	0.763	0.763

Panel (b): Cost Performance

	(1)	(2)	(3)	(4)	(5)
$\hat{\theta}_{public}$	0.018	0.016	0.039	0.039	0.045
	(0.009)	(0.010)	(0.011)	(0.011)	(0.010)
$\hat{\theta}_{private}$	-0.003	0.004	-0.028	-0.030	-0.037
	(0.013)	(0.014)	(0.016)	(0.016)	(0.015)
N	98,089	98,089	98,089	98,089	98,089
R^2	0.036	0.120	0.184	0.189	0.206
$Avg_{services}$	0.832	0.832	0.832	0.832	0.832
Avg_{works}	0.918	0.918	0.918	0.918	0.918
Controls		✓	✓	✓	✓
Bureau Fixed Effects			✓	✓	✓
State Fixed Effects				✓	✓
Object Fixed Effects					✓

Notes: results of the DDD regression of *Time Performance* - panel (a) - and *Cost Performance* - panel (b) - on public oversight and private oversight treatment indicators. Column 1 reports the results of a plain DDD regression, specification 2 adds controls for contract and bureau characteristics. Columns 3-5 include an increasing number of fixed effects (bureau, year, state, and contract object). To deal with a collection of minor problems about normality, heteroscedasticity or observations that exhibit large residuals, leverage or influence, standard errors are estimated using the Eicker-Huber-White estimators. In each panel, $Avg_{services}$ and Avg_{works} account for the average outcome in the Services and the Public Works treatment group, respectively. Only coefficients related to treatment effects are reported. For the complete set of regressions coefficients please refer to Tables (A.9) and (A.10).

this implies that firms in the treatment group are less likely to revise contract terms and, when they do, the overrun is less on both time and cost dimensions; instead, waiving private oversight does not affect the likelihood of revision and has an opposite effect at the intensive margin for cost and time dimensions.

Table 4: Triple Difference - Contract Outcomes

	Time		Cost	
	(1)	(2)	(3)	(4)
	Extensive	Intensive	Extensive	Intensive
$\hat{\theta}_{public}$	-0.052	-57.585	-0.030	-42.187
	(0.014)	(8.502)	(0.018)	(11.097)
$\hat{\theta}_{private}$	0.038	73.544	0.017	-67.807
	(0.024)	(24.612)	(0.041)	(21.274)
N	98,089	45,795	98,089	53,147

Notes: *Time* - columns (1)-(2) - and *Cost* - columns (3)-(4) - extensive and intensive margin analysis. For each outcome, the odd column reports the results of a linear probability model of *Time/Cost Amended* on public oversight and private oversight treatment indicators plus controls for firm size, tender type, whether the firm is a limited liability company, whether the contract was signed during the last week of the fiscal year, bureau, year, state, and contract object fixed effects. Even olumns report the estimates of a DDD with the same set of controls on the *Average Time/Cost Overrun*.

V.3 Robustness Checks

We test the robustness of our baseline DDD findings over three dimensions. First, we check whether the results are robust across different subsamples of contracts; second, we test the suitability of the DDD empirical strategy in our framework; finally, we rule out the possibility that our estimates of $\theta_{private}$ are driven by differential shocks in the construction industry.

Testing the sample selection We start the analysis by checking whether the estimated parameters are robust to changes in the estimation sample. We are concerned, on the one hand, that big contracts in the upper control group may drive the results and, on the other hand, that there exist possible sources of contamination due to the unobserved engineers' value: since we do not observe the *ex-ante* valuation of the project we may misclassify part of the projects to the treatment/control groups.[55] In Table (5), in columns 2 and 3 we

[55] As remarked above, we only observe an indicator variable for the valuation being above or below the threshold.

report the *range* and *sanitary* models, which restrict the sample to all contracts valued less that \$500,000 and exclude observations in a 10 percent window around the \$100,000 and \$150,000 thresholds, respectively. In column 4, we take care of possible outliers in terms of outcomes and exclude contracts that are associated to overruns worth nine times the expected contract values (performance lower than 0.1) in terms of cost and time, separately. In the appendix, we perform the same exercise when cost and time overrun exceed four times the expected outcomes (i.e., performances < 0.25). In column 5 we show the results obtained by running model (3) with the lower control group (i.e., all contracts valued between \$50,000 and \$100,000).

Testing the empirical design The feasibility of the DDD analysis relies on a series of assumptions that should be carefully analyzed. In particular, one should ensure that the average outcome of the control and the treatment group hold parallel before the intervention date, as we show in Figure 1 for the present case in a way that the estimated treatment effect should be zero before the treatment date. Following Autor (2003), we test this assumption by running a model according to the following specification

$$y_{ist} = \gamma_s + \delta_t + \sum_{j=-m}^{q} \theta_{public,j} Waiver_{it+j} + Constr_{it} \tag{4}$$

$$+ \sum_{j=-m}^{q} \theta_{private,j} (Waiver * Constr)_{it+j} + X_{ist} + \epsilon_{ist} \tag{5}$$

where γ_s and δ_t are Government bureau and fiscal year fixed effects, respectively, X_{ist} are contract-level controls and the specification includes $m = 5$ leads and $q = 4$ lags. A graphical representation of the results is reported in Figure (2): we excluded $t = -1$ (i.e., the dummy relative to the FY 2010) in order to avoid collinearity issues and $t^{\star}$ stands for year $= 2011$. Although the results - mostly for time performance - appear to be noisy, there is no clear evidence of non-parallel pre-trends and the public and private oversight effects reflect the estimated parameters.

Table 5: Triple Difference - Sample Selection Robustness Checks

Panel (a): Time Performance

	(1) Base	(2) Range	(3) Sanitary	(4) Hperf	(5) Lower
$\hat{\theta}_{public}$	0.048	0.051	0.053	0.128	0.147
	(0.011)	(0.011)	(0.013)	(0.023)	(0.026)
$\hat{\theta}_{private}$	-0.067	-0.064	-0.068	-0.035	-0.027
	(0.027)	(0.027)	(0.026)	(0.037)	(0.037)
N	98,089	65,971	81,848	21,218	22,055
R^2	0.110	0.111	0.108	0.124	0.128
$Avg_{services}$	0.805	0.805	0.805	0.833	0.802
Avg_{works}	0.811	0.811	0.811	0.821	0.828

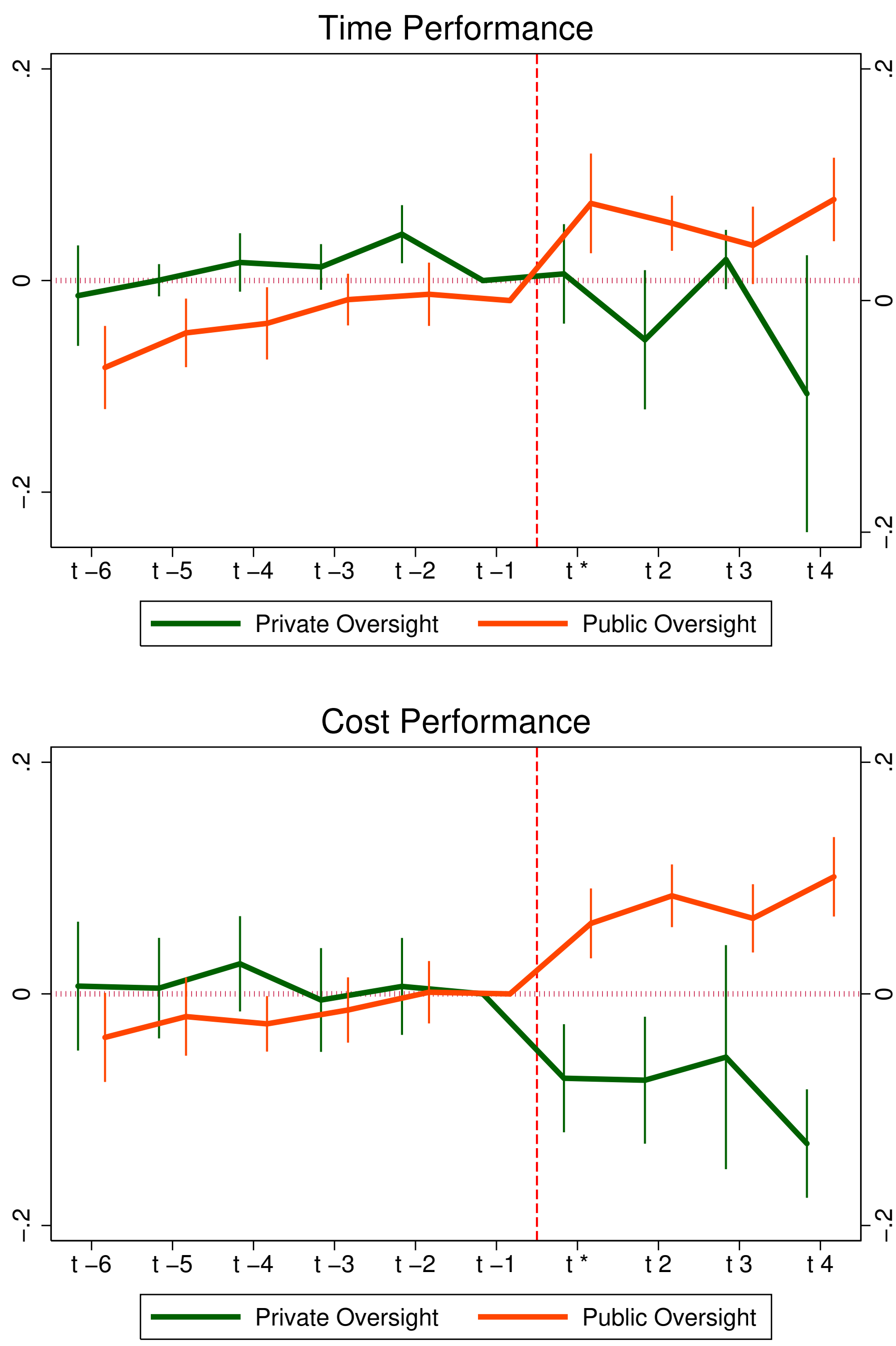
Panel (b): Cost Performance

	(1)	(2)	(3)	(4)	(5)
$\hat{\theta}_{public}$	0.045	0.044	0.046	0.121	0.151
	(0.010)	(0.009)	(0.011)	(0.024)	(0.027)
$\hat{\theta}_{private}$	-0.037	-0.048	-0.038	-0.051	-0.055
	(0.015)	(0.014)	(0.017)	(0.016)	(0.018)
N	98,089	65,971	81,848	21,455	22,055
R^2	0.206	0.202	0.207	0.168	0.202
$Avg_{services}$	0.881	0.881	0.881	0.887	0.880
Avg_{works}	0.945	0.945	0.945	0.945	0.948

Notes: results of various DDD models of *Time* - panel (a) - and *Cost* - panel (b) - dimensions on public and private oversight treatment indicators plus controls for firm size, tender type, whether the firm is a limited liability company, whether the contract was signed during the last week of the fiscal year, bureau, year, state, and contract object fixed effects. *Base* reports the baseline model results; *Range* reports results of the above model run on a sample of contracts trimmed at a face value of $500,000; in *Sanitary* we account for the possible contamination at the $100,000 and $150,000 thresholds dropping all contracts whose face value lies within a 10 percent band around; *HPerf* indicates the exclusion of contracts with performance lower than 0.1 *Lower* is run with the *lower control group*. In each panel, $Avg_{services}$ and Avg_{works} account for the average outcome in the Service and the Public Works treatment group, respectively.

Another major concern when using the DD design is the possibility that estimates suffer from the effects of confounding - and unobservable - factors changing over time. In this sense, our estimates of public oversight may suffer from selection problem, as we consider all contracts above (or below) the threshold as a unique control group. Given the heterogenous nature of contracts within the control groups - they either range $150,000-695,000,000 or

Figure 2: $\hat{\theta}_{public}$ and $\hat{\theta}_{private}$ - Time and Cost performance



Notes: Leads and lags analysis of DDD for private and public oversight. We exclude $t-1$ from the analysis. Time (above panel) and cost performance (below panel).

$50,000-100,000 - and the panel dimension of our sample, it is possible to construct a finer pool of controls through the synthetic control (SC) group approach, developed by Abadie, Diamond and Hainmueller (2010), to remove arbitrariety in control group definition. In order to do that, we first group upper and lower control groups together and partition them in 68 groups according to the contract amount, then we build the SC group as the convex combination of partitions that most closely resemble the treatment group in terms of pre-2011 outcomes and predictors.[56] In Figure (3) we plot the treatment group (solid line) and SC group (dotted line) yearly series for *cost* and *time performance*. The gross average difference between treatment and SC groups are very close to the baseline DDD estimates for public oversight, amounting to 0.053 (6.4 percent) and 0.046 (6 percent) for *cost performance* and *time performance*, respectively. In order to evaluate the reliability of SC estimates, and to test the sensitivity of the results to the choice of the optimal SC group, we run the inference analysis proposed by Abadie, Diamond and Hainmueller (2010) by checking whether we would obtain the same results if we choose a group at random among the partitions. The complete series of placebo tests is presented in Figure (4) alongside the actual treatment group results: the exercise strongly supports the existence of a significant public oversight treatment effect.

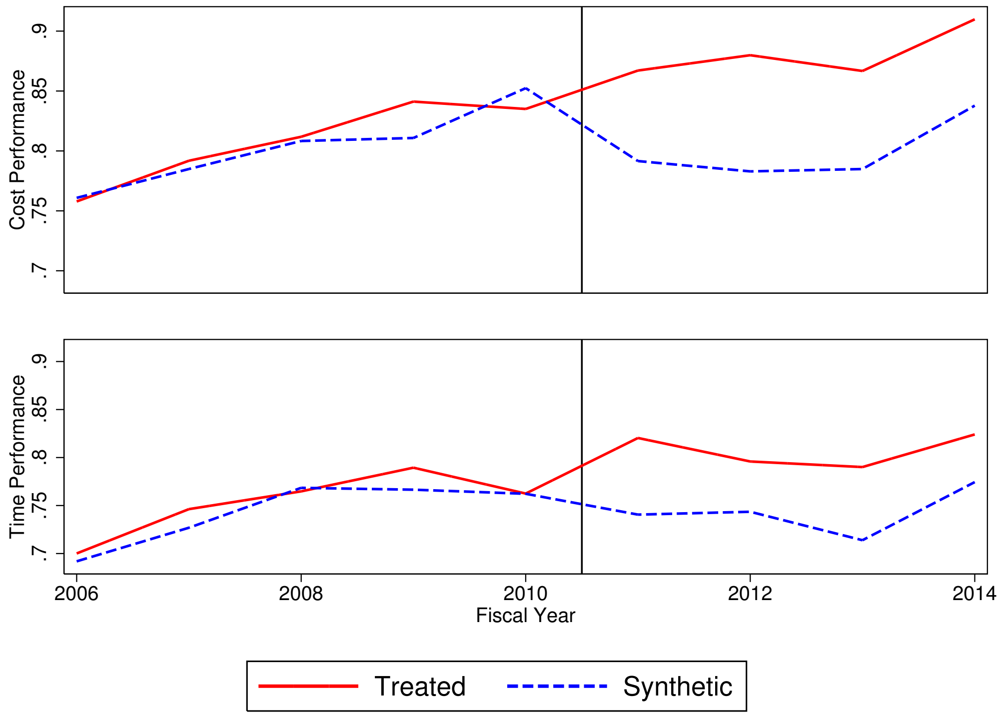
Testing the "Differential Shock Hypothesis" Another concern in our setting is the possibility that a shock hit the construction industry in FY2011 - and lasting for a long period of time - and that the estimate of $\theta_{private}$ is thus capturing the time- and cost-dimension effects on contract outcomes of such differential shock instead of an actual treatment effect.

In order to address such concern, we start by considering input prices that are typical of construction industry and account for related observable shocks.[57] In figure (5) we report the monthly time series of the average prices of the main input for construction projects - i.e., Cement, Insulation Materials, Asphalt Felts Coatings, Concrete, Construction Sand,

[56]SC builds upon the setting of the standard DD model, but makes two changes as i) it allows for time-varying individual-specific heterogeneity and ii) it takes a data driven approach to forming counterfactuals, i.e. selecting the control group. Rather than picking one of these two large control groups at a time, SC creates an optimal weighted average of contracts above \$150,000 and below \$100,000. Each of the selected group is associated to the respective weight. In the appendix - Figure (A.7) - we report all groups associated with non-zero weights, and their amount.

[57]The underlying assumption being that, if really unexpected, an input price shock may have dramatic effects on renegotiations both in terms of time and costs.

Figure 3: Synthetic Control Group - Parallel Trends



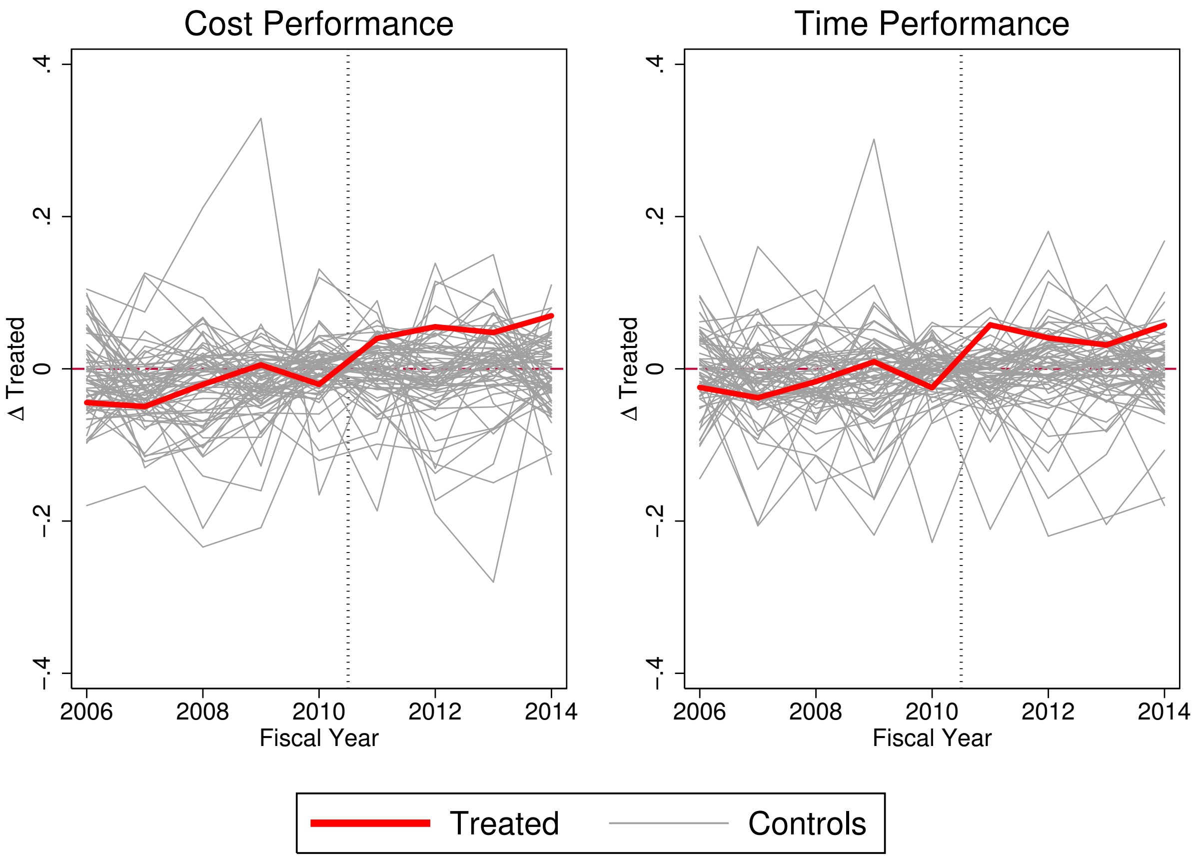
Notes: Parallel trends in SC group analysis for public oversight - *cost* (above panel) and *time performance* (below panel). The solid line is the treatment group, while the dotted line represents the SC group and is the weighted average of 68 partitions of upper and lower control groups.

and ready-mix Concrete - ranging 2005-2016.[58] Although quite stable on average, prices follow different dynamics while the price of *Asphalt Felts Coatings* shows a clear jump in 2008, and kept being pretty noisy since then. In a related exercise, we run the model (3) adding the full set of prices as additional controls. Estimated parameters, reported in Table (6) - columns 2 and 5 - do not change with respect to the baseline results, showing a striking robustness: in turn, this indicates that either input prices do not alter the likelyhood of contractors to renegotiate ex-post, or that bidders include future input prices cost into their bids ex-ante.

The second step in order to rule out the possibility that our estimates pick up specific differential shocks is testing whether our results are due to firms self-sorting in particular

[58] Prices are reported *per unit*. The unit is a standard measure and changes with the input considered (e.g., the concrete's standard unit is the cubic meter, while for the insulation materials is the square meter).

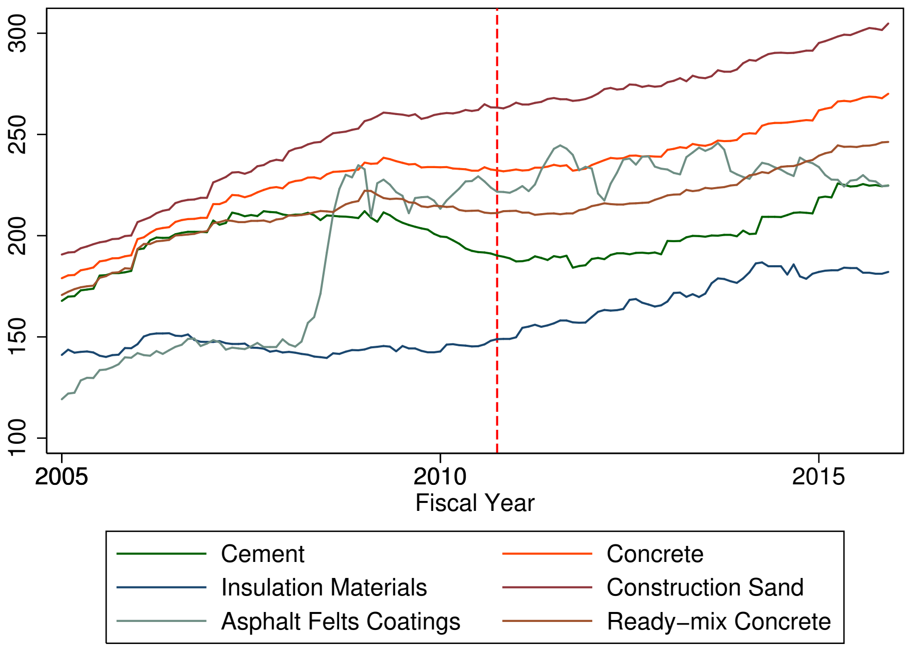
Figure 4: Synthetic Control Group - Inferential analysis



Notes: Inference in synthetic control group analysis for public oversight - cost (left panel) and time performance (right panel). Estimated gaps in yearly-average *cost performance* and *time performance* between the treatement group and all control groups. The idea of Abadie, Diamond and Hainmueller (2010) is extending the idea of a placebo study to produce quantitative inference in comparative case studies. The idea of the placebo test mimics the classic framework for permutation inference, where the distribution of a test statistic is computed under random permutations of the sample units' assignments to the intervention and nonintervention groups. As in permutation tests, we apply the SC method to every potential control in our donor sample. This allow to assess whether the effect estimated by the SC design for the group affected by the intervention is large relative to the effect estimated for a group chosen at random.

markets (i.e., selection on observables). Hence, we follow a vast literature (see Abadie (2005) and DiNardo, Fortin and Lemieux (1996), among the others) in implementing a propensity score (triple) differences estimator. Thanks to the richness of our data, we first estimate the propensity score $e(\mathbf{z})$, defined as the probability for a contract i of being in the treatment group ($Waiver = 1$) conditional on a set of observables $\mathbf{z}$ unrelated to the outcomes of

Figure 5: Input Prices Time Series - Constructions



Notes: monthly time series of the construction input prices averaged over the US, 2005-2016. The red vertical dashed line represents October, 2010.

interest.[59] We then make use of the score to generate the weights $w_i = Waiver_i + (1 - Waiver_i)\frac{e(\mathbf{z}_i)}{1-e_i(\mathbf{z})}$ as in Hirano and Imbens (2001), and proceed by estimating a weighted version of equation (3). The estimated parameters account both for the transitory shocks thanks to the balanced control/treatment groups and for the structural heterogeneity among groups, which is differenced out by the difference-in-differences approach. In columns 3 and 6 of Table (6) we report the relative estimated parameters for time and cost performance, respectively; while the sample is slightly reduced due to the lack of matches on observables for some treated contracts, results appear to be extremely robust to the selection issue.

[59]See Appendix for additional details on the model, variables used, and balancing test across treatment/control groups.

Table 6: Triple Difference - Differential Shocks Robustness Checks

	Time Performance			Cost Performance		
	(1)	(2)	(3)	(4)	(5)	(6)
	Base	Input	PSM	Base	Input	PSM
$\hat{\theta}_{public}$	0.048	0.483	0.041	0.045	0.000	0.056
	(0.011)	(0.084)	(0.013)	(0.010)	(0.076)	(0.011)
$\hat{\theta}_{private}$	-0.067	0.045	-0.059	-0.037	-0.032	-0.052
	(0.027)	(0.055)	(0.043)	(0.015)	(0.020)	(0.022)
N	98,089	22,055	96,077	98,089	22,055	97,345
R^2	0.110	0.130	0.164	0.206	0.203	0.237
$Avg_{services}$	0.763	0.763	0.763	0.837	0.837	0.837
Avg_{works}	0.763	0.763	0.763	0.921	0.921	0.921

Notes: results of various DDD models of *Time Performance* - columns 1-3 - and *Cost Performance* - columns 4-6 - on public and private oversight treatment indicators plus controls for firm size, office size, tender type, whether the contract format is a fixed price and bureau, year, state, and contract object fixed effects. Columns 1 and 4 report baseline estimates. In columns 2 and 5 we add controls for construction input prices, while columns 3 and 6 report estimates of the model weighted by the propensity score.

VI Discussion

Performance Bonding: Moral Hazard vs. Adverse Selection With the triple difference approach we are able to identify the net effect of performance bonding on contract outcomes. However, the estimates alone do not help distinguish whether this is due to adverse selection, to moral hazard, or to both. In order to test hypotheses *b.1)* and *b.2)*, we partition firms in the construction treatment subgroup into three clusters: *Stayers* - those firms that win at least one contract both before and after October, 2010 within the treatment band; *Exiters* that win at least one contract within the band before, but are not awarded a contract worth less than $150,000 afterwards; and *Entrants*, which are never awarded a contract worth more than $100,000 before the threshold revision, but win at least one contract in the treatment band afterwards. Within the latter two clusters, we can further identify two subclusters: *Segment Entrants* (*Segment Exiters*) are those firms that enter (exit) the treatment band after being awarded at least one contract below (above) $100,000 during the pre-treatment period.

According to CGH (hypothesis *b.1*), the pool of winning firms must change with the treatment. This would be reflected in a moderate share of *Stayers* and a high number of both *Exiters* and *Entrants*. On top of that, the share of limited liability firms should be higher in the pool of entrants, given that their aggressive bidding strategy is not counterbalanced by higher surety premia.[60] Conversely, hypothesis *b.2)* does not entail any structural adjustment in the set of winning firms: according to this hypothesis sureties do not exert any ex-ante screening and affect the outcomes only through the ex-post supervision of work progress. In this scenario, the same firms are awarded contracts both before and after the reform but their performance is negatively affected by the absence of the surety, whose monitoring role is crucial to avoid misbehavior. Hence, we would expect to observe a relatively high number of *Stayers* and a moderate or null turnover; moreover, *Entrants* and *Exiters* should be similar in size and observable characteristics.

In Table (7), panel *(a)* reports summary statistics of each group within the PB treatment group. The treated sample spans 2008-2013 with the treatment threshold fixed at October, 2010. In order to enhance comparability we repeat the exercise (panels *b* and *c*) on two placebo subsamples spanning 5 years but not containing the threshold revision date - i.e., all contracts before and all contracts after October, 2010. We place two placebo thresholds in 2008 and 2012, respectively. *Stayers* constitute 6.8 percent of the sample in the treatment group - *Panel (a)* - while *Entrants* and *Exiters* account for 31.2 percent and 62 percent, respectively. The *Stayers* figure is similar and remains low in all subsamples, this being an indicator of a high turnover level in the market, but the proportion of exiters to entrants is reversed in both panels *(b)* and *(c)* with respect to the treatment group.[61] This indicates that an unusually high number of firms exit the federal construction market following the 2011 reform - three times as many as those in the Placebo 1 subsample and almost 7 times the number of *Exiters* in the Placebo 2 subsample - but are not replaced by a comparable

[60]Limited liability firms are able to bid more aggressively because their financial responsibility in case of loss is bounded. For more details on this, see CGH.

[61]See, for example, the Bureau of Labor Statistics on turnover levels at http://www.bls.gov/iag/tgs/iag23.htm.

number of firms.[62]

An interesting piece of evidence in favor of the presence of adverse selection is provided by the *Segment Entrants* and the *Segment Exiters* figures. The share of exiters is remarkably regular across the three panels, lingering below the 5 percent.[63] On the other hand, the volume of *Segment Entrants* varies substantially, starting at similar levels before the reform and increasing dramatically during treatment period (15 percent), only to decrease again (9 percent) afterwards. This is consistent with hypothesis *b.1)*: a negligible share of firms step up from below the $100,000 threshold to the treatment group when sureties screen bidders ex ante, while the transition is relatively easy in their absence. We document an analogous pattern in the share of entrant limited liability companies, which peaks during the treatment period and decreases afterwards. Following CGH, we interpret this as a signal of adverse selection.

Public Monitoring: Competence and the Red Tape Effect In public procurement, quality of buyers is a crucial feature to account for. Competence in providing goods and services helps in selecting the best contractors, properly designing contracts and avoiding misconduct. Moreover, high-quality officers might be able to minimize the red tape effect of bureaucracy through rapid and efficient monitoring activities. Thus, we would expect the public oversight treatment effect to be particularly intense in contracting offices with low levels of competence and, conversely, contracts managed by expert officials not to be affected. Using the same data, Decarolis et al. (2018) already showed the importance of assessing the contracting office's competence in dealing with public procurement performance. In order to do so, they propose a measure of office quality defined as the persistency of contract performance within the same purchasing organization and for the same category of procured good or service over time. More specifically, for each contract awarded, they measure the weighted average outcome (in terms of past cost performance and past time performance,

[62] Weighting the comparison for the ratio of total number of firms in the samples yields 2.45 and 4.57 times more, respectively.

[63] We obtain similar results for segment Exiters with different placebo subsamples, where we define placebo treatments according to different levels of contract value thresholds. Results are available from the authors upon request.

Table 7: Summary statistics: Stayers, Entrants and Exiters in the Construction Market

Panel (a): Treated - 2011 Threshold

	Stayers			Entrants			Exiters		
	Mean	Median	N	Mean	Median	N	Mean	Median	N
No. of Employees	5,868.09	41.4	37	3,119.84	18	165	158.17	12	336
Annual Revenue	1,443.38	12.1	37	293.29	3.40	165	62.09	2.17	336
Limited Company	0.16	0	37	0.21	0	165	0.10	0	336
Past Time Perf e	0.79	0.80	37	0.80	0.80	161	0.82	0.83	334
Past Cost Perf e	0.90	0.92	37	0.87	0.89	161	0.89	0.90	334
Segment Entrants				0.15	0	165			
Segment Exiters							0.04	0	336

Panel (b): Placebo 1 - 2008 Threshold

	Stayers			Entrants			Exiters		
	Mean	Median	N	Mean	Median	N	Mean	Median	N
No. of Employees	7,164.09	11.9	26	225.48	10.5	307	1,321.22	25	113
Annual Revenue	1,365.78	5.71	26	74.77	2.15	307	102.08	3	113
Limited Company	0.08	0	26	0.12	0	307	0.03	0	113
Past Time Perf e	0.78	0.75	26	0.82	0.84	306	0.81	0.81	109
Past Cost Perf e	0.91	0.93	26	0.89	0.90	306	0.91	0.93	109
Segment Entrants				0.05	0	307			
Segment Exiters							0.04	0	113

Panel (c): Placebo 2 - 2012 Threshold

	Stayers			Entrants			Exiters		
	Mean	Median	N	Mean	Median	N	Mean	Median	N
No. of Employees	2,256.63	34.1	19	3,828.86	19	144	455.11	30	51
Annual Revenue	684.95	11.5	19	484.14	4.35	144	79.47	4.50	51
Limited Company	0.26	0	19	0.19	0	144	0.18	0	51
Past Time Perf e	0.77	0.75	19	0.80	0.80	142	0.80	0.81	48
Past Cost Perf e	0.92	0.93	19	0.86	0.86	142	0.90	0.93	48
Segment Entrants				0.10	0	144			
Segment Exiters							0.04	0	51

Notes: descriptive statistics relative to the group of *Stayers*, *Entrants* and *Exiters*. These have been defined according to the actual threshold revision in 2011 - panel (a) - or with two placebo reform dates in 2008 - panel (b) - and 2012 - panel (c).*No. of Employees* is the average number of employees over the time span considered; *Annual Revenue* reports the last 3 years' average revenue - in US$ thousands ; *Limited Company* indicates a limited liability company; *Past Time* and *Past Cost Performance* are defined as in Decarolis et al. (2018). *Segment Entrants* indicates the share of Entrant firms which only appeared below the threshold prior to October, 2010, while *Segment Exiters* reports the number of Exiter firms which appear only above the threshold before the reform.

Table 8: Triple Difference - High- and Low-type contracting offices

Panel (a): Performance Measures

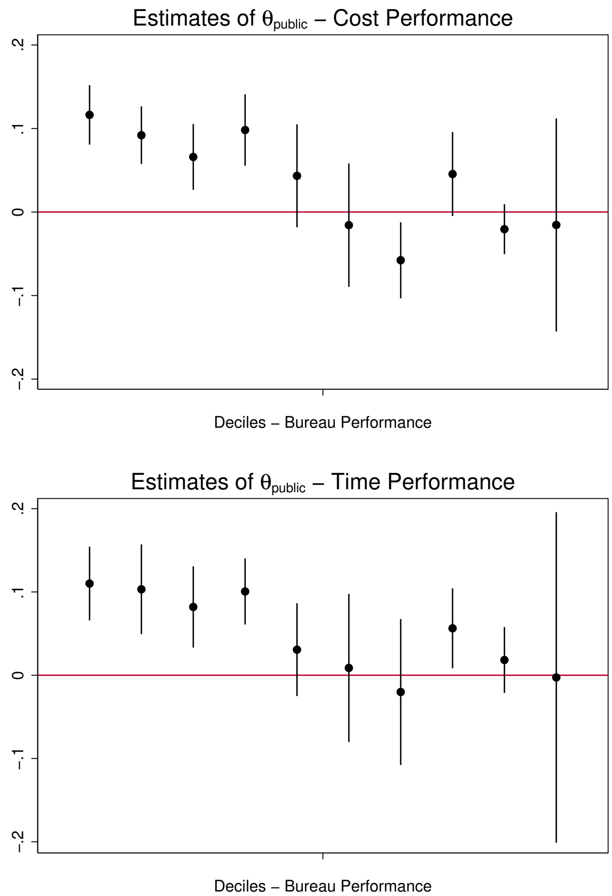
	Time Performance		Cost Performance	
	(1) H	(2) L	(3) H	(4) L
$\hat{\theta}_{public}$	0.020	0.084	0.000	0.077
	(0.014)	(0.014)	(0.012)	(0.013)
N	50,087	48,002	47,720	50,369
R^2	0.096	0.062	0.149	0.155
$Avg_{services}$	0.772	0.754	0.847	0.817
Avg_{works}	0.766	0.757	0.914	0.927

Panel (b): Renegotiation Measures

	Time Overrun		Cost Overrun	
	H	L	H	L
$\hat{\theta}_{public}$	-28.263	-108.375	36.606	-48.232
	(19.769)	(19.260)	(37.929)	(45.539)
N	50,087	48,002	47,720	50,369
R^2	0.097	0.120	0.157	0.156
$Avg_{services}$	188.348	181.850	158.467	235.468
Avg_{works}	1.000	149.418	59.080	45.425

Notes: DDD regressions of *Time Performance* and *Cost Performance* - panel (a) - and *Time Overrun* and *Cost Overrun* - panel (b) - on public and private oversight treatment indicators plus controls for firm size, tender type, whether the firm is a limited liability company, whether the contract was signed during the last week of the fiscal year, bureau, year, state, and contract object fixed effects. Odd columns report results for contracts managed by high-competence contracting offices, even columns refer to low-competence. Standard errors are in parentheses. In each panel, $Avg_{services}$ and Avg_{works} account for the average outcome in the Services and the Public Works treatment group, respectively.

Figure 6: $\hat{\theta}_{public}$ on deciles of bureau performance



Notes: DDD estimated βs on deciles of bureau performance for cost (left panel) and time performance (right panel).

separately) of the contracts awarded by the same bureau in the past. A similar rationale applies to our framework, and we build the same measure of past performance for each contract i as $past\,performance_{ik} = \frac{\sum_{j_k < i_k} w_{j_k} * performance_{j_k}}{\sum_{j_k} w_{j_k}}$, where k stands for the contracting office, $performance_{j_k}$ is the performance value achieved in the j^{th} contract awarded by k and w_{j_k} are weights for smoothing past observations.

We repeat the exercise for cost and time performance and, aggregating *past performance* at bureau level, we exploit the distribution of the variable to partition the sample according to the contracting office competence level and run the baseline DDD estimation on the relative subsamples. In Figure (6) we plot the estimated θ_{public} *per* decile of the competence distribution, with the relative standard error. It is immediate to note how - consistently with the red tape effect prediction - we estimate a positive and significant effect of public oversight exemption only for contracting offices in the lowest deciles of the distribution, while we cannot find any effect for more competent bureaus. On a similar page, in Table (8) we report the results of the estimation obtained partitioning the sample in *High* (H)- and *Low* (L)-type bureaus, depending on the median value of past performances. Not surprisingly, we find that the detrimental effect of red tape is large and significant for contracts awarded by the least competent contracting offices only. Hence, public oversight worsens contractors' performance if and only if the procurement process is managed by a low-quality bureau, while the scores for offices with a track record of excellence in project management are not altered after the treatment.

VII Conclusions

The oversight of public procurement process is among the most keenly debated tools for policy-makers and practitioners due to its heterogeneous nature: it aims at reducing the asymmetry of information between public buyers and sellers, but it can be a source of waste and inefficiency. Our findings provide evidence of both effects for low-value contracts: on the one hand, the public oversight overloads the contractors with unnecessary red tape, on the other hand the private oversight, by requiring the presence of surety companies into the procurement process, leads to the selection of better contractors and to advances in overall performances. Our result support the thesis that public bodies are not always necessary for the performance of a task carried out in the public interest and that, accordingly, outsourcing to private players would generate important benefits for the society.

From an economic perspective, the interplay between the public buyer and the surety companies poses several interesting questions, which our results help to address. First, we confirm empirically the results of the theoretical model proposed by CGH: the screening exerted by the performance bonding is effective in lowering the likelihood of a low-quality winners in public procurement tenders. Our results suggest that surety companies are helpful in gathering “soft” information on firm types - elements like managerial practices, effort exerted in previous projects, etc., typically impossible to be considered for public players - and turn it into verifiable information, i.e., the premium passed through to the bids. In this sense, the presence of private firms lowers the asymmetry of information involved in the procurement process. Second, the public buyer, by indirectly paying the bond premium, ensures the contract is completed either by the original contractor or by the intervention of the surety itself, in case the seller faces issues endangering its ability to perform the project. This makes the performance bond a perfectly suited tool for procurement practitioners, as the final completion of the project is the first-order concern for procurers. Our findings suggest indeed that the performance bonding has effects on the extensive margin of renegotiations, too, thus highlighting the prominent role played by surety companies during the operational phase.

Our findings on the negative effects of public oversight - even though in accordance with previous literature - are especially relevant when framed into the context of bureacratic competence. We find that the magnitude of red tape hinges on the persistent incapacity in project management by public officials and we show the substantial heterogeneity of these effects with respect to the office competence level, in a way that contracts awarded by high-quality government offices do not seem to be affected by the cutoff revision, which is instead extremely beneficial for contracts procured by low-quality procurers. This is in line with the literature highlighting the role of competence in public offices as the first-order concern for contractual performances - e.g., Bandiera, Prat and Valletti (2009) for the Italian case.

In addition, our paper provides straightforward policy recommendations in the context of the US federal procurement: for low-valued procurement contracts we show that it would be possible to keep the benefits of the private oversight while avoiding the negative effects of the public oversight by moving the thresholds for the application of the Miller Act and for the Simplified Acquisitions independently. Moreover, the presence and the role of surety companies are not alien from potential agency issues: there exists the possibility that very big contractors “capture” small sureties and that, on the other hand, insurers and insured collude to raise prices at the expenses of procurers. Unfortunately, FPDS data miss crucial information on the surety companies backing up contractors: among the most interesting extensions to the present work would be a dynamic analysis of the surety market in order to address some of the above issues.

References

Abadie, Alberto. 2005. "Semiparametric difference-in-differences estimators." *The Review of Economic Studies*, 72(1): 1–19.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association*, 105: 493–505.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review*, 93(1): 113–132.

Athey, S., J. Levin, and E. Seira. 2011. "Comparing Open and Sealed-Bid Auctions: Evidence from Timber Auctions." *The Quarterly Journal of Economics*, 126: 207–257.

Autor, David H. 2003. "Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing." *Journal of labor economics*, 21(1): 1–42.

Bajari, Patrick, and Gregory Lewis. 2014. "Moral Hazard, Incentive Contracts, and Risk: Evidence from Procurement." *The Review of Economic Studies.*

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

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

Bajari, P., S. Houghton, and S. Tadelis. 2014. "Bidding for Incomplete Contracts: An Empirical Analysis of Adaptation Costs." *American Economic Review*, 104(4): 1288–1319.

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, R Hanna, J Kyle, B Olken, and S Sumarto. 2017. "The role of competition in effective outsourcing: Subsidized food distribution in Indonesia." *Journal of Political Economy.*

Banerjee, Abhijit V., and Esther Duflo. 2000. "Reputation effects and the limits of contracting: a study of the Indian Software Industry." *The Quarterly Journal of Economics*, 115(3): 989–1017.

Berck, P., and S.B. Villas-Boas. 2016. "A note on the triple difference in economic models." *Applied Economics Letters.*

Bergman, A. M., P. Johansson, S. Lundberg, and G. Spagnolo. 2016. "Privatization and quality: Evidence from elderly care in Sweden." *Journal of Health Economics*, 49: 109–119.

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.

Board, Simon. 2007. "Bidding into the Red: A Model of Post-Auction Bankruptcy." *Journal of Finance*, 62(6): 2695–2723.

Bozeman, Barry. 1993. "A Theory Of Government "Red Tape"." *Journal of Public Administration Research and Theory*, 3(3): 273–303.

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

Calveras, A., J.J. Ganuza, and E. Hauk. 2004. "Wild bids. Gambling for resurrection in procurement contracts." *Journal of Regulatory Economics*, 26(1): 41–68.

Calvo, E., R. Cui, and J.C. Serpa. 2016. "Oversight and Efficiency in Public Projects: A Regression Discontinuity Analysis." *Working Paper.*

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

Cerulli, Giovanni. 2016. *Econometric evaluation of socio-economic programs: theory and applications.* Springer.

Coviello, Decio, and Mario Mariniello. 2014. "Publicity Requirements in Public Procurement: Evidence from a Regression Discontinuity Design." *Journal of Public Economics*, Forthcoming.

Coviello, Decio, Andrea Guglielmo, and Giancarlo Spagnolo. 2017. "The Effect of Discretion on Procurement Performance." *Management Science*, 63(5).

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

Decarolis, Francesco, Leonardo M. Giuffrida, Elisabetta Iossa, Vincenzo Mollisi, and Giancarlo Spagnolo. 2018. "Bureaucratic Competence and Procurement Outcomes." National Bureau of Economic Research Working Paper 24201.

Decarolis, F., R. Pacini, and G. Spagnolo. 2016. "Past Performance and Procurement Outcomes." *Working Paper.*

DiNardo, John, Nicole M Fortin, and Thomas Lemieux. 1996. "Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach." *Econometrica.*

Di Tella, R., and E. Schargrodsky. 2003. "The Role of Wages and Auditing during a Crackdown on Corruption in the City of Buenos Aires." *Journal of Law & Economics*, 46(1): 269–292.

Ferraz, C., and F. Finan. 2008. "Exposing corrupt politicians: the effects of Brazil's publicly released audits on electoral outcomes." *The Quarterly Journal of Economics*, 123(2): 703–745.

Ferraz, C., and F. Finan. 2011. "Motivating politicians: The impacts of monetary incentives on quality and performance." *Working Paper.*

Fisman, R., and E. Miguel. 2007. "Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets." *Journal of Political Economy*, 115(6).

Fisman, R., and R. Gatti. 2006. "Bargaining for bribes: The role of institutions." *CEPR Discussion Papers*.

GAO, US. 2013. "Agencies need to strengthen oversight of multibillion dollar investments in operations and maintenance." United States Government Accountability Office GAO-14-66.

Garvin, M., K. Molenaar, D. Navarro, and G. Proctor. 2011. "Key performance indicators in public-private partnerships." US Department of Transportation FHWA-PL-10-029.

Gransberg, D. D., P. Kraft, and H. Park. 2014. "Performance Bond: Cost, Benefit, and Paradox for the Public Highway 3 Agencies." *Transportation Research Record*, 2408: 3–9.

Guasch, J. L., J.J. Laffont, and S. Straub. 2008. "Renegotiation of concession contracts in Latin America: Evidence from the water and transport sectors." *International Journal of Industrial Organization*, 22(7): 1267–1294.

Hart, O., A. Shleifer, and R. W. Vishny. 1997. "The Proper Scope of Government: Theory and an Application to Prisons." *Quarterly Journal of Economics*, 112(4): 1127–1161.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica*, 66(5): 1017–1098.

Hirano, Keisuke, and Guido W Imbens. 2001. "Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization." *Health Services and Outcomes research methodology*, 2(3): 259–278.

Hyytinen, Ari, Sofia Lundberg, and Otto Toivanen. 2018. "Design of public procurement auctions: evidence from cleaning contracts." *The RAND Journal of Economics*.

Kaufman, H. 1977. *Red Tape: Its origins, uses, and abuses.* Brookings Institution Press.

Kelman, Steven. 1990. *Procurement and Public Management: The Fear of Discretion and the Quality of Government Performance.* University Press of America Lanham.

Kelman, Steven. 2005. *Unleashing Change: A Study of Organizational Renewal in Government.* Brookings Institution Press.

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

Laffont, Jean-Jacques, and Jean Tirole. 1990. "Adverse selection and renegotiation in procurement." *The Review of Economic Studies*, 57(4): 597–625.

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. 2017. "Procurement Contracting with Time Incentives: Theory and Evidence." *The Quarterly Journal of Economics*, forthcoming.

Liebman, Jeffrey B., and Neale Mahoney. 2017. "Do Expiring Budgets Lead to Wasteful Year-End Spending? Evidence from Federal Procurement." *American Economic Review*, 107(11): 3510–49.

Marion, Justin. 2007. "Are Bid Preferences Benign? The Effect of Small Business Subsidies in Highway Procurement Auctions." *Journal of Public Economics*, 91(7-8): 1591–1624.

Marion, Justin. 2009. "How Costly Is Affirmative Action? Government Contracting and California's Proposition 209." *The Review of Economics and Statistics*, 91(3): 503–522.

McCrary, J. 2008. "Manipulation of the running variable in the regression discontinuity design: A density test." *Journal of Econometrics.*

Olken, Benjamin A. 2006. "Corruption and the costs of redistribution: Micro evidence from Indonesia." *Journal of Public Economics.*

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

Palguta, J., and F. Pertold. forthcoming. "Manipulation of Procurement Contracts: Evidence from the Introduction of Discretionary Thresholds." *American Economic Journal: Economic Policy.*

Reinikka, R., and J. Svensson. 2004. "Local Capture: Evidence from a Central Government Transfer Program in Uganda." *The Quarterly Journal of Economics.*

Rosembaum, Paul R., and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70(1): 41–55.

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

Shleifer, A., and R. W. Vishny. 1998. *The Grabbing Hand: Government Pathologies and Their Cures.* Harvard University Press.

Spiller, Pablo T. 2008. "An institutional theory of public contracts: regulatory implications." *NBER Working Paper*, , (14152): 1–33.

Szucs, Ferenc. 2018. "Discretion and Corruption in Public Procurement." *Working Paper.*

A Appendix - For Online Publication

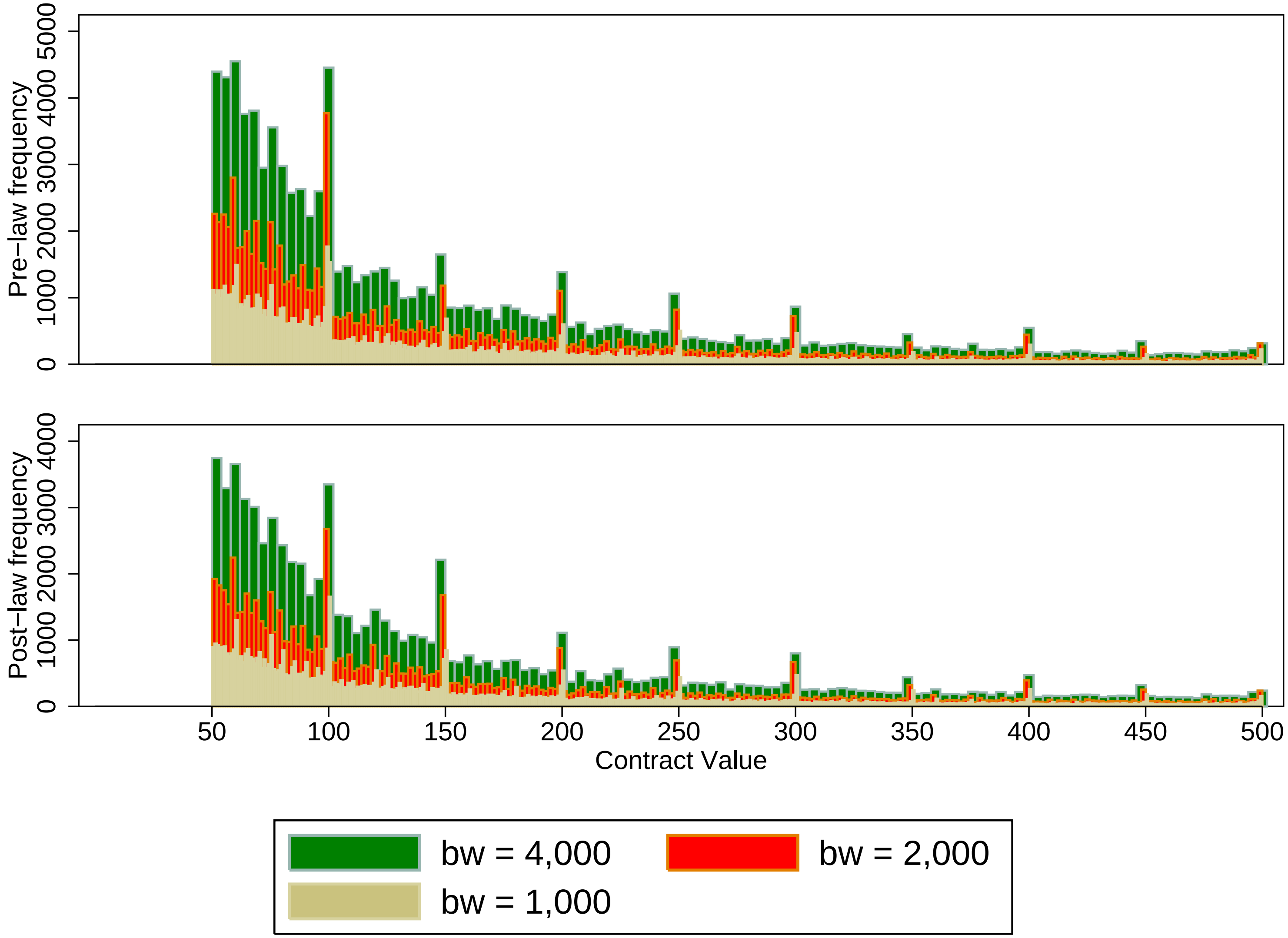
A.1 Data features

The previous literature used legislative thresholds in RDD in order to analyze whether contract-allocation procedures with different levels of discretion affect procurement outcomes (see Coviello and Mariniello (2014) and Coviello, Guglielmo and Spagnolo (2017)). As shown in Figure (A.1), most of the "round" values - i.e., multiples of $50,000 - show significant jumps in frequency both in pre- and post-law update: this is a clear indication of sorting before different procedural thresholds. Some of these cutoffs, such as the one at $200,000, seem to be mainly psychological, since FAR provides no legislative provision attached to these contracts values. As a result, due to this likely endogenous sorting of contracts both prior to and after October, 2010, the performance of firms above and below the threshold may have been different even before the reform. Had we ignored this and performed an RDD analysis, we could have retrieved severely upward biased estimates due to a pre-existent discontinuity.[64]

In order to test for endogenous sorting or discontinuities in the forcing variable, we performed the McCrary (2008) density test for post-law data for both construction (Figure A.2) and services (Figure A.3). The sharp discontinuity of the running variable at the $150,000 threshold, highlighted by the graph and confirmed by the significant test results, rules out any possibility of running a usual RDD with our data. The endogenous sorting and its increment after the update of the SAT threshold are clear-cut evidence of the facts that (i) winning firms' incentives to sort themselves below $150,000 became stronger, (ii) the effect of the confounding policy discontinuity on the potential outcome is not constant over time, and (iii) this effect was the same in the pre- and post-treatment period (as Figure (A.4), displaying the pre- and post-reform contract value density around the $150,000 threshold, displays). This is confirmed by Figure (A.5), showing that pre-reform contract value density around the $100,000 threshold is higher than post-reform.

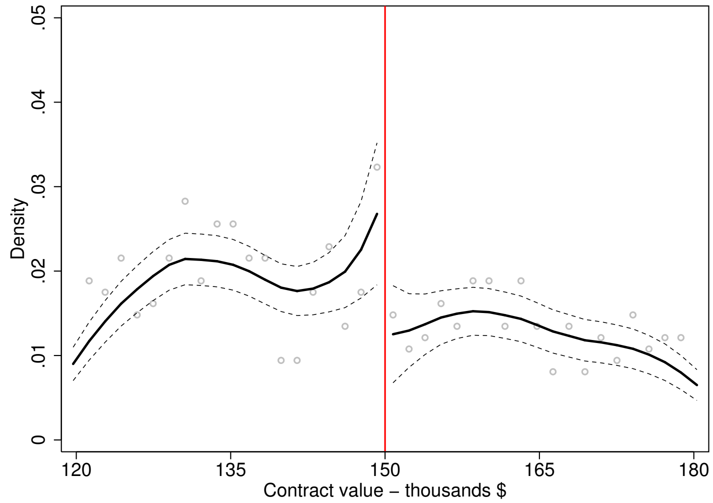
[64] An instance of bunching in procurements just below legislative thresholds is presented in Palguta and Pertold (forthcoming)

Figure A.1: Contract frequency: various binsizes



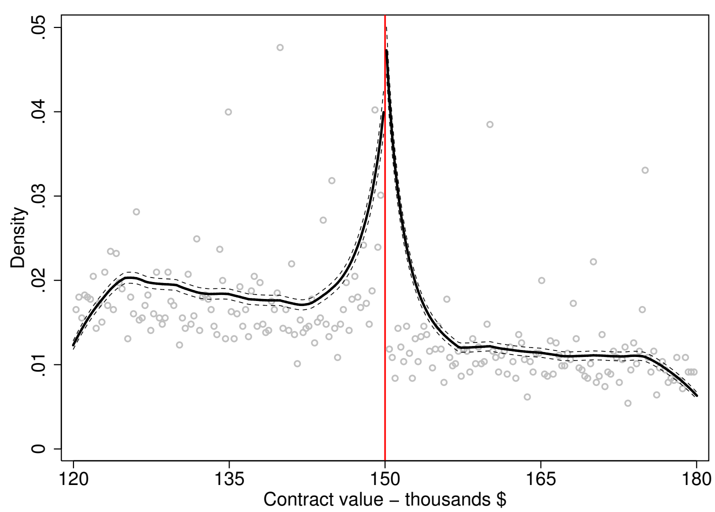
Notes: Contract frequencies pre- and post-reform, absolute values. The histogram is plotted for different bin widths: $4,000 (green), $2,000 (red) and $1,000 (yellow).

Figure A.2: McCrary Density Test - Construction Contracts



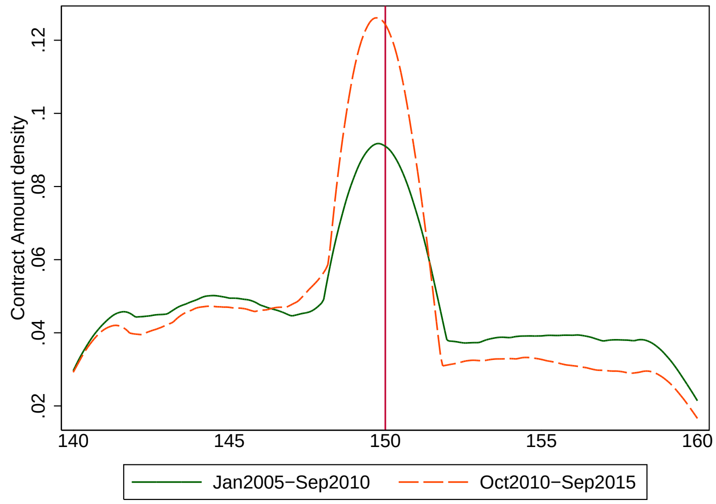
Notes: Mccrary density in a +/- \$30,0000 window around the \$150,000 threshold. The dots represent the density of projects in different intervals of project budget, the solid line represent a kernel estimate of the density, and the two dashed lines are 95 percent confidence intervals. Construction contracts only.

Figure A.3: McCrary Density Test - Full Sample



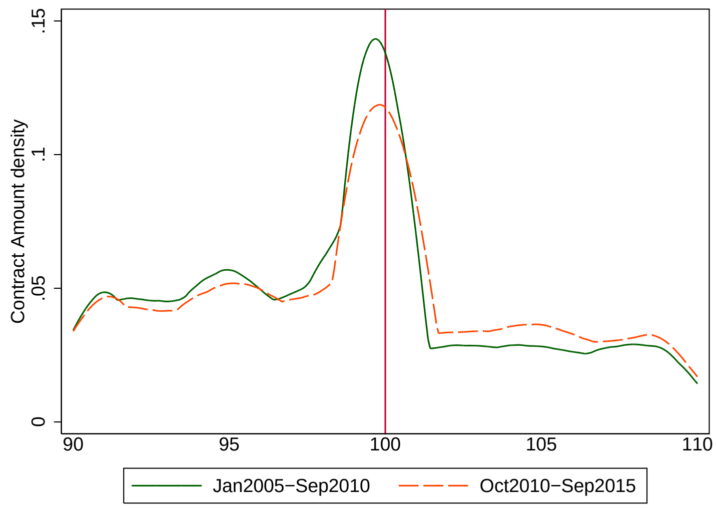
Notes: McCrary density in a +/- \$30,000 window around the \$150,000 threshold. The dots represent the density of projects in different intervals of project budget, the solid line represent a kernel estimate of the density, and the two dashed lines are 95 percent confidence intervals.

Figure A.4: Pre- and Post-reform contract value density



Notes: Contract amount density in a +/- $20,000 window around the $150,000 threshold before (solid) and after (dotted) - the threshold revision.

Figure A.5: Pre- and Post-reform contract value density - 100,000



Notes: Contract amount density in a +/- $20,000 window around the $100,000 threshold before (solid) and after (dotted) the threshold revision.

A.2 Summary statistics

Table (A.1) reports selected summary statistics for the lower control group of services - along with those for the treatment group, already presented in (2) - before and after the October, 2010 reform. In Table (A.2) we report summary statistics for the constructions sample and distinguish, again, the above treatment/control before/after groups scheme. Tables (A.3) and (A.4) display the number of contracts by department and year before and after the reform when considering the upper and the lower control group, respectively, for the analysis. Finally, Table (A.6) presents the composition of treatment and both lower and upper control groups in terms of the categories of service content of contracts representing around 97 per cent of observations. The distribution of categories is quite regular across groups. This exercise is aimed to show that the treatment group truly constitutes a representative subsample in our population of contracts.

Table A.1: Summary statistics - Services Treatment Group with Lower Control Group

	Lower Control Group							
	Before				After			
	Mean	SD	Median	N	Mean	SD	Median	N
Time Performance	0.7	0.3	0.74	51,245	0.7	0.3	0.82	52,737
Num Time Amendments	2.4	2.8	1	51,245	1.4	2.1	0	52,737
Prob Time Revision	0.6	0.5	1	51,245	0.6	0.5	1	52,737
Avg Time Overrun	199.4	252.3	128.6	27,887	214.0	287.9	115.6	22,684
Cost Performance	0.7	0.3	0.91	51,245	0.8	0.3	0.98	52,737
Num Cost Amendments	2.8	3.2	2	51,245	1.8	2.6	1	52,737
Prob Cost Revision	0.6	0.5	1	51,245	0.5	0.5	1	52,737
Avg Cost Overrun	273.8	333.5	205.1	31,587	307.8	413.7	176.8	27,481
Contract Value	1,896.8	9,311.5	421.7	51,245	1,098.6	4,334.6	300	52,737
# Contractual Days	417.6	404.5	364	51,245	305.1	222.1	357	52,737
Offers received	5.4	18.7	2	51,245	5.7	18.7	2	52,737

	Treatment Group							
	Before				After			
	Mean	SD	Median	N	Mean	SD	Median	N
Time Performance	0.8	0.3	1	13,774	0.8	0.3	1	2,061
Num Time Amendments	1.1	2.0	0	13,774	0.7	1.5	0	2,061
Prob Time Revision	0.4	0.5	0	13,774	0.3	0.5	0	2,061
Avg Time Overrun	229.9	287.3	141	4,763	187.6	250.5	100	565
Cost Performance	0.8	0.3	1	13,774	0.9	0.2	1	2,061
Num Cost Amendments	1.3	2.3	0	13,774	0.8	1.7	0	2,061
Prob Cost Revision	0.4	0.5	0	13,774	0.3	0.5	0	2,061
Avg Cost Overrun	101.2	125.9	64.2	5,377	67.5	96.7	34.1	625
Contract Value	122.2	15.2	121.2	13,774	122.0	16.1	120.1	2,061
# Contractual Days	298.9	342.8	215	13,774	228.2	191.1	205	2,061
Offers received	4.7	14.5	2	13,774	3.9	18.9	1	2,061

Notes: the table reports descriptive statistics for both the Lower Control Group (upper panel) and the treatment group (lower panel), before (left side) and after (right side) the threshold revision. *Time* and *Cost Performance* are relative measures of performance - bounded 0 to 1; *Num Time* and *Num Cost Amendments* count the number of amendments per contract, while the relative *Prob* is a binary variable that takes value 1 in case of any amendment occurs; *Avg Time* and *Avg Cost Overrun* account for the average extra time or extra cost and is defined only for contracts which had at least one amendment; *Contract Value* is expressed in US$ thousands; *Offers Received* report the number of offers received per tender.

Table A.2: Summary statistics - Public Works Sample with Upper Control Group

	Upper Control Group							
	Before				After			
	Mean	SD	Median	N	Mean	SD	Median	N
Time Performance	0.7	0.3	0.67	5,787	0.7	0.3	0.76	2,789
Num Time Amendments	3.2	3.0	2	5,787	2.6	2.9	1	2,789
Prob Time Revision	0.7	0.4	1	5,787	0.7	0.5	1	2,789
Avg Time Overrun	113.1	155.8	69	4,045	100.4	147.5	56.1	1,746
Cost Performance	0.9	0.2	0.94	5,787	0.9	0.2	0.96	2,789
Num Cost Amendments	3.6	3.4	2	5,787	2.8	3.2	1	2,789
Prob Cost Revision	0.8	0.4	1	5,787	0.7	0.5	1	2,789
Avg Cost Overrun	123.5	186.7	47.9	4,466	190.0	357.2	66.7	1,876
Contract Value	6,265.4	20,015.3	661.0	5,787	4,510.8	12,132.9	484.7	2,789
# Contractual Days	335.4	283.9	249	5,787	277.1	220.7	205	2,789
Offers received	4.3	4.0	3	5,787	4.4	4.6	3	2,789

	Treatment Group							
	Before				After			
	Mean	SD	Median	N	Mean	SD	Median	N
Time Performance	0.7	0.3	0.95	947	0.8	0.3	1	90
Num Time Amendments	1.1	1.7	0	947	0.5	0.9	0	90
Prob Time Revision	0.5	0.5	1	947	0.4	0.5	0	90
Avg Time Overrun	131.8	169.0	83.8	417	133.4	129.2	105	27
Cost Performance	0.9	0.2	1	947	0.9	0.1	1	90
Num Cost Amendments	1.0	1.6	0	947	0.5	0.9	0	90
Prob Cost Revision	0.5	0.5	0	947	0.3	0.5	0	90
Avg Cost Overrun	39.1	99.9	10.8	446	30.7	36.2	17.3	26
Contract Value	123.7	14.7	123.4	947	123.8	14.0	123.6	90
# Contractual Days	171.4	210.5	115	947	182.2	186.9	121	90
Offers received	3.7	11.4	3	947	2.7	1.8	3	90

Notes: the table reports descriptive statistics for both the Public Works Upper Control Group (upper panel) and the Public Works treatment group (lower panel), before (left side) and after (right side) the threshold revision. *Time* and *Cost Performance* are relative measures of performance - bounded 0 to 1; *Num Time* and *Num Cost Amendments* count the number of amendments per contract, while the relative *Prob* is a binary variable that takes value 1 in case of any amendment occurs; *Avg Time* and *Avg Cost Overrun* account for the average extra time or extra cost and is defined only for contracts which had at least one amendment; *Contract Value* is expressed in US$ thousands; *Offers Received* report the number of offers received per tender.

Table A.3: Number of contracts by department and year - Upper control group and treatment group

	Pre-Treatment						Post-Treatment				
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Dept of Agriculture	76	42	58	57	75	153	145	118	85	110	12
Dept of Commerce	635	639	605	660	948	1,187	1,073	862	857	666	121
Dept of Defense	336	380	404	502	330	384	368	441	307	265	26
Dept of Education	12	17	18	19	27	37	15	26	10	7	1
Dept of Energy	8	13	24	17	19	48	34	38	23	20	5
Dept of Health and Human Services	132	324	462	971	989	1,422	717	147	21	10	3
Dept of Homeland Security	23	20	17	13	13	24	13	14	8	17	3
Dept of Housing	12	13	17	24	34	137	100	112	86	94	17
Dept of Interior	462	469	497	398	656	1,118	525	482	361	411	146
Dept of Justice	278	286	394	332	365	532	481	387	369	392	44
Dept of Labor	1,262	3,218	4,792	4,862	4,807	4,649	3,821	3,716	2,398	2,104	302
Dept of State	41	60	59	65	57	44	50	41	35	36	8
Dept of Treasury	109	101	77	106	105	127	88	104	76	105	15
Dept of Transportation	1,209	1,156	1,036	991	1,095	1,205	1,160	1,035	899	844	105
Dept of Veteran Affairs	843	846	1,197	1,206	1,170	1,097	1,248	1,249	1,029	887	191
Environmental Protection Agency	47	48	55	44	49	111	48	18	23	12	1
General Services Administration	410	789	736	737	1,047	1,602	1,116	963	896	902	115
NASA	357	374	698	679	991	1,469	1,812	2,000	2,010	1,993	582
Nuclear Regulatory Commission	144	45	74	67	78	165	148	142	143	129	26
National Science Foundation	22	107	129	187	247	379	405	382	302	252	76
Office of Personnel Management	306	400	442	331	346	397	362	418	291	269	39
Small Business Administration	213	437	726	907	840	1,050	1,048	1,019	836	656	127
Social Security Administration	749	596	787	1,227	1,963	3,354	3,223	2,972	2,823	2,325	958
N	7,686	10,380	13,304	14,402	16,251	20,691	18,000	16,686	13,888	12,506	2,923

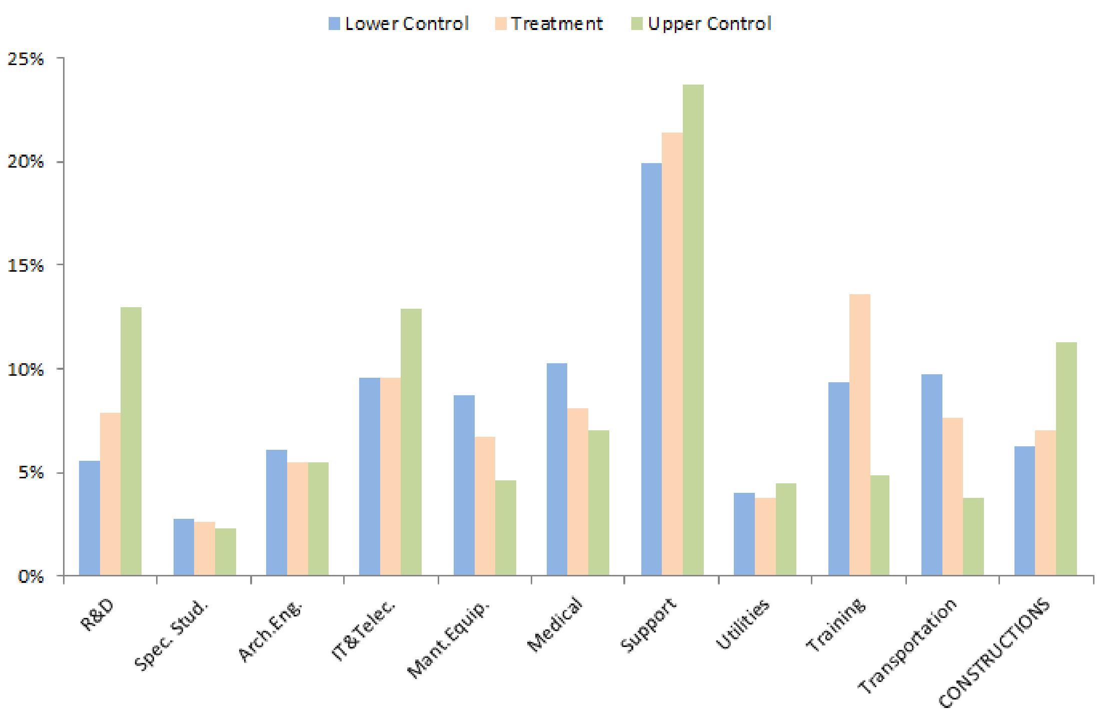
Notes: Number of contracts by year/department; Upper control group and treatment group. Pre-treatment period: January 2005 to October, 2010; post-treatment periods: October 2010 to September 2015.

Table A.4: Number of contracts by agency and year - Lower control group and treatment group

	Pre-Treatment						Post-Treatment				
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Dept of Agriculture	82	70	73	60	124	159	224	157	124	153	17
Dept of Commerce	440	426	367	439	517	706	665	558	581	518	149
Dept of Defense	231	250	293	299	232	209	201	222	222	195	34
Dept of Education	5	8	10	16	17	43	22	27	12	10	2
Dept of Energy	1	5	8	8	15	28	12	24	19	21	4
Dept of Health and Human Services	443	597	851	1,470	1,476	1,314	434	101	31	16	5
Dept of Homeland Security	21	15	23	13	22	30	12	15	12	10	3
Dept of Housing	8	18	23	26	26	265	182	178	125	83	10
Dept of Interior	422	507	530	492	738	1,037	587	619	401	432	348
Dept of Justice	380	295	301	367	381	443	413	371	316	289	43
Dept of Labor	551	1,914	3,441	3,451	3,342	3,364	2,890	2,716	1,885	1,780	356
Dept of State	11	15	21	19	14	30	21	16	14	21	3
Dept of Treasury	93	97	56	49	53	56	83	71	61	69	14
Dept of Transportation	890	933	909	903	953	837	724	716	616	622	146
Dept of Veteran Affairs	542	467	596	594	551	477	601	598	509	485	156
Environmental Protection Agency	18	31	48	30	30	38	9	8	9	8	
General Services Administration	304	759	853	980	1,117	1,323	1,020	861	770	768	149
NASA	193	240	395	452	781	1,100	1,500	1,431	1,399	1,790	656
Nuclear Regulatory Commission	100	46	58	56	58	104	105	83	70	80	19
National Science Foundation	24	101	133	165	209	361	301	270	203	161	68
Office of Personnel Management	190	200	229	204	243	252	261	194	147	126	28
Small Business Administration	77	213	549	527	490	716	693	549	504	356	65
Social Security Administration	573	586	673	981	1,728	2,897	2,757	2,215	1,933	1,670	968
N	5,599	7,793	10,440	11,601	13,117	15,789	13,717	12,000	9,963	9,663	3,243

Notes: Number of contracts by year/department; Lower control group and treatment group. Pre-treatment period: January 2005 to October, 2010; post-treatment periods: October 2010 to September 2015.

Figure A.6: Categories of services by Treatment Group and Lower/Upper Control Groups



Notes: Number of contracts by year/department; Upper control group and treatment group. Pre-treatment period: January 2005 to October, 2010; post-treatment periods: October 2010 to September 2015.

A.3 Legal Details - public oversight

The FAR requires procurement officers to exert an ex-ante and ex-post

The sealed bidding acquisition process involves six steps in the US Government procurement: i) *performance work statement*: officers draft a written specification for the agency requirements; ii) *solicitation preparation*: the invitation for bid has to describe the requirements clearly, accurately, and completely without unnecessarily restrictive specifications that would unduly limit the number of bidders; iii) *invitation for bid*: the officers publicize the tender on Federal Business Opportunities, in case of domestic award, or local newspapers or websites, in case of abroad award, and in sufficient time to enable prospective bidders to prepare and submit bids; iv) bid submission: bidders submit sealed bids to be publicly opened at the time and place stated in the invitation for bid; v) bid evaluation: the officers evaluate bids without discussion with bidders; and vi) contract award: the contracting officer awards to the lowest-priced, responsive bidder. Differently, federal procurement negotiation refers to any acquisition method that is not sealed bidding and involves nine steps. i) performance work statement: as in the case of sealed bidding, officers draft a written specification for the agency requirements; ii) request for proposals preparation: as in the case of an invitation for bid, the contracting agency describes its requirements clearly and accurately and does so without unnecessarily restrictive requirements which would unduly limit the number of offerors; iii) request for proposal: the contracting officers publicizes the request for proposal through all appropriate means, including Federal Business Opportunities, local newspapers, embassy, consulate or procurement office website, or any other locally acceptable means of advertising the requirement; iv) proposal submission: offerors prepare and submit technical and price proposals in response to the request for proposals; v) proposals evaluation: a technical evaluation panel evaluates the technical proposals against the technical evaluation criteria included in the request for proposal. The contracting officer evaluates the price proposals to determine whether the proposed prices are reasonable and reflect the offeror's understanding of, and ability to perform, the contract. The contracting officer then determines which proposals are in the competitive range based on both technical and price factors. vi) Discussions conducted. The contracting officer holds oral and/or written discussions with each offeror in the competitive range to resolve uncertainties and to provide each offeror with an understanding of the technical or price weaknesses in its proposal; vii) final proposal revisions request: once discussions are concluded, the contracting officer issues a written request for final proposal revisions from all offerors in the competitive range; viii) final proposal revisions evaluation: the technical evaluation panel and the contracting officer evaluate final proposal revisions in the same manner as initial offers; and ix) contract award: the contracting officer awards a contract to the responsible offeror whose offer is most advantageous, considering price and other related evaluation factors as stated in the solicitation.

Furthermore, according to FAR 42.1104a) the contract administration office determines the extent of production surveillance on the basis of 1) the degree of importance to the Government assigned by the contracting officer to the supplies or services; and 2) Consideration of the following factors: (i) Contract requirements for reporting production progress and performance; ii) The contract performance schedule; iii) The contractor's production plan; iv) The contractor's history of contract performance; v) The contractor's experience with the contract supplies or services; vi) The contractor's financial capability; vii) Any supplementary written instructions from the contracting office.

Operational Stage Auditing officers within each contracting office are in charge to require sellers to (i) complete expenditure justification forms and submit cost or pricing data certifying that expenses are based on adequate price competition; and (ii) submit reports on the project's progress to specific evaluation teams.[65] In the first case, according to FAR 15.402, in establishing the reasonableness of the offered prices, the contracting officer shall obtain i) certified cost or pricing data, along with data other than certified cost or pricing data as necessary to establish a fair and reasonable price and ii) the type and quantity of data necessary to establish a fair and reasonable price, but not more data than is necessary. The second case involves Government review and analysis of (i) contractor performance plans, schedules, controls, and industrial processes; and ii) the contractor's actual performance under them.

[65]The number and type of checks are similar for each contracting office, as provided by the FAR, and their scope is analogous; we can coherently group them into one set.

A.4 Additional Robustness Checks

We test the robustness of our baseline DDD findings on three dimensions. We first check in Table (A.5) whether the results are robust in different subsets of the baseline sample; secondly, we run the estimation on absolute overrun values, namely *Cost Overrun* and *Time Overrun*, in Table (A.6); finally, we show the robustness of our results using the lower control group in Table (A.7).

We are concerned with possible sources of contamination due to the unobserved engineers' estimate value: in fact, we do not observe the *ex-ante* valuation of the project but only whether it lays within the \$100,000-150,000 band (i.e., if it is subject to the SAP or not. In order to test whether the possible misclassification of contracts to the treatment group drives our results, in Table (A.5), columns (2) and (6), we rule out observations in a (-\$10,000, + \$10,000) window around \$100,000 - *Contamin*. Results are robust and, if anything, there is a positive variation in the magnitude of the estimates. In columns (3) and (7) we run the baseline model on a narrower time window (2008-2013), reducing the risk that our results are driven by long-term trends in the data.

Public officers are specifically demanded to carry out an ex-post monitoring of work status in all services contracts except constructions. In the construction industry, instead, this is a task for the surety. We want to check whether the presence of an additional service-only specific treatment drives our results - mainly with respect to the *public oversight* estimates. Hence, we modify the triple difference model in equation (3) to:

$$\begin{aligned} y_{ijt} = \alpha &+ \beta_1 Waiver_{it} + \beta_2 Post_{it} + \theta_{public} \left(Waiver_{it} * Post_{it} \right) + \\ &+ \beta_3 Constr_{it} + \beta_4 \left(Constr_{it} * Waiver_{it} \right) + \beta_5 \left(Constr_{it} * Post_{it} \right) + \\ &+ \theta_{private} \left(Constr_{it} * Waiver_{it} * Post_{it} \right) + \theta_{pc} T^{pc}_{it} + \gamma X_{it} + \delta_t + \zeta_j + \varepsilon_{ijt} \end{aligned} \tag{6}$$

where T^{pc}_{it} is a binary treatment variable active for services-only contracts above the threshold. Results of the estimation are reported in columns (4) and (8): parameter $\hat{\theta}_{pc}$

Table A.5: Triple Difference - Differential Shocks Robustness Checks

	Time Performance				Cost Performance			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Base	Contamin	2008-2013	Public Check	Base	Contamin	2008-2013	Public Check
$\hat{\theta}_{public}$	0.048	0.052	0.048	0.048	0.045	0.046	0.050	0.042
	(0.011)	(0.012)	(0.010)	(0.014)	(0.010)	(0.010)	(0.010)	(0.012)
$\hat{\theta}_{private}$	-0.067	-0.061	-0.058	-0.067	-0.037	-0.036	-0.037	-0.034
	(0.027)	(0.023)	(0.035)	(0.029)	(0.015)	(0.016)	(0.018)	(0.017)
N	98,089	92,101	68,251	98,089	98,089	92,101	68,251	98,089
R^2	0.110	0.109	0.119	0.110	0.206	0.206	0.211	0.206
$Avg_{services}$	0.763	0.760	0.751	0.763	0.832	0.827	0.818	0.832
Avg_{works}	0.763	0.764	0.752	0.763	0.918	0.914	0.913	0.918

Notes: results of various DDD models of *Time Performance* - columns (1) to (4) - and *Cost Performance* - columns (5) to (8) - on public and private oversight treatment indicators plus controls for firm size, tender type, whether the firm is a limited liability company, whether the contract was signed during the last week of the fiscal year, bureau, year, state, and contract object fixed effects. *Base* reports the baseline model results; in *Contamin* we account for the possible contamination at the $100,000 threshold dropping all contracts whose face value lies between $90,000 and $110,000; *2008-13* is self-explanatory and *Public Check* reports results of a modified model with a specific treatment dummy for public on-site monitoring. In each panel, $Avg_{services}$ and Avg_{works} account for the average outcome in the Service and the Public Works treatment group, respectively. Standard errors are clustered at the object * year level.

(unreported) is not statistically significant.

A second set of checks are reported in Table (A.6), in which the baseline DDD model, with an increasing number of controls and fixed effects, is implemented on different dependent variables: panel (a) reports results on *Time Overrun* (in days), while panel (b) reports *Cost Overrun* (in dollars).

There are two contract groups potentially suitable for use as a control. Although in the baseline regressions we use contracts always subject to oversight (upper control group), we also have data on all those contracts that are never subject to oversight (lower control group). Table (A.7) presents the baseline and robustness triple difference regressions using the lower control group. The estimates prove to be robust to the change in control group both in terms of sign and magnitude.

Table A.6: Triple Difference - Overruns

Panel (a): Time Overrun

	(1)	(2)	(3)	(4)	(5)
$\hat{\theta}_{public}$	-45.135	-40.632	-66.450	-66.258	-62.684
	(11.741)	(14.021)	(14.515)	(14.515)	(14.446)
$\hat{\theta}_{private}$	48.670	42.410	122.463	125.616	121.099
	(19.777)	(20.623)	(23.868)	(24.176)	(23.355)
N	98,089	98,089	98,089	98,089	98,089
R^2	0.031	0.049	0.114	0.117	0.126
$Avg_{services}$	185.228	185.228	185.228	185.228	185.228
Avg_{works}	117.565	117.565	117.565	117.565	117.565

Panel (b): Cost Overrun

	(1)	(2)	(3)	(4)	(5)
$\hat{\theta}_{public}$	13.407	44.724	-11.120	-10.778	-16.752
	(24.366)	(29.639)	(31.046)	(30.235)	(29.185)
$\hat{\theta}_{private}$	-239.847	-248.666	-193.147	-190.790	-206.209
	(90.362)	(89.775)	(70.828)	(62.860)	(65.324)
N	98,089	98,089	98,089	98,089	98,089
R^2	0.046	0.121	0.181	0.187	0.201
$Avg_{services}$	259.193	259.193	259.193	259.193	259.193
Avg_{works}	72.954	72.954	72.954	72.954	72.954
Controls		✓	✓	✓	✓
Bureau Fixed Effects			✓	✓	✓
State Fixed Effects				✓	✓
Object Fixed Effects					✓

Notes: results of the DDD regression of *Time Overrun* - panel (a) - and *Cost Overrun* - panel (b) - on public oversight and private oversight treatment indicators plus increasing controls and fixed effects. In each panel, $Avg_{services}$ and Avg_{works} account for the average outcome in the Services and the Public Works treatment group, respectively. Standard errors are clustered at the object * year level.

Table A.7: Triple Difference - Lower Control Group

Panel (a): Time Performance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Base	Contamin	Sanitary	2008-2013	Range	Public Check	Input	PSW
$\hat{\theta}_{public}$	0.047	0.041	0.035	0.052	0.064	0.035	0.047	0.079
	(0.009)	(0.009)	(0.010)	(0.009)	(0.011)	(0.010)	(0.009)	(0.011)
$\hat{\theta}_{private}$	-0.005	0.009	0.013	-0.000	-0.024	0.007	-0.003	-0.007
	(0.022)	(0.025)	(0.026)	(0.020)	(0.025)	(0.022)	(0.021)	(0.042)
N	74,973	60,459	54,854	52,092	26,540	74,973	74,973	73,856
R^2	0.118	0.118	0.119	0.133	0.125	0.119	0.120	0.134
$Avg_{services}$	0.802	0.802	0.800	0.803	0.802	0.802	0.802	0.802
Avg_{works}	0.828	0.826	0.824	0.830	0.834	0.828	0.828	0.828

Panel (b): Cost Performance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\hat{\theta}_{public}$	0.059	0.052	0.049	0.064	0.076	0.046	0.058	0.085
	(0.006)	(0.006)	(0.007)	(0.007)	(0.009)	(0.007)	(0.006)	(0.009)
$\hat{\theta}_{private}$	-0.044	-0.039	-0.036	-0.046	-0.056	-0.032	-0.043	-0.065
	(0.012)	(0.015)	(0.015)	(0.014)	(0.015)	(0.012)	(0.011)	(0.016)
N	74,973	60,459	54,854	52,092	26,540	74,973	74,973	75,373
R^2	0.172	0.169	0.171	0.190	0.181	0.173	0.174	0.208
$Avg_{services}$	0.880	0.880	0.877	0.875	0.879	0.880	0.880	0.883
Avg_{works}	0.948	0.947	0.947	0.947	0.944	0.948	0.948	0.949

Notes: results of various DDD models of *Time Performance* - panel (a) - and *Cost Performance* - panel (b) - on public and private oversight treatment indicators plus controls for firm size, tender type, whether the firm is a limited liability company, whether the contract was signed during the last week of the fiscal year, bureau, year, state, and contract object fixed effects. *Base* reports the baseline model results; in *Contamin* we account for the possible contamination at the $100,000 threshold dropping all contracts whose face value lies between $90,000 and $110,000; in a similar fashion *Sanitary* model deals with the contamination at both thresholds ($100,000 and $150,000) by dropping two 10 percent sanitary bands around; *2008-13* is self-explanatory and *Range* reports results of the same model applied to a sample of contracts trimmed at a face value of $500,000. *Public Check* - abbreviated to *P.Check* - reports results of a modified model with a specific treatment dummy for public on-site monitoring while in the *Input* model we added controls for public works input prices throughout the period. Finally, *PSW* reports the results of the Propensity Score weighted DDD. In each panel, $Avg_{services}$ and Avg_{works} account for the average outcome in the Service and the Public Works treatment group, respectively. Standard errors are clustered at the object * year level.

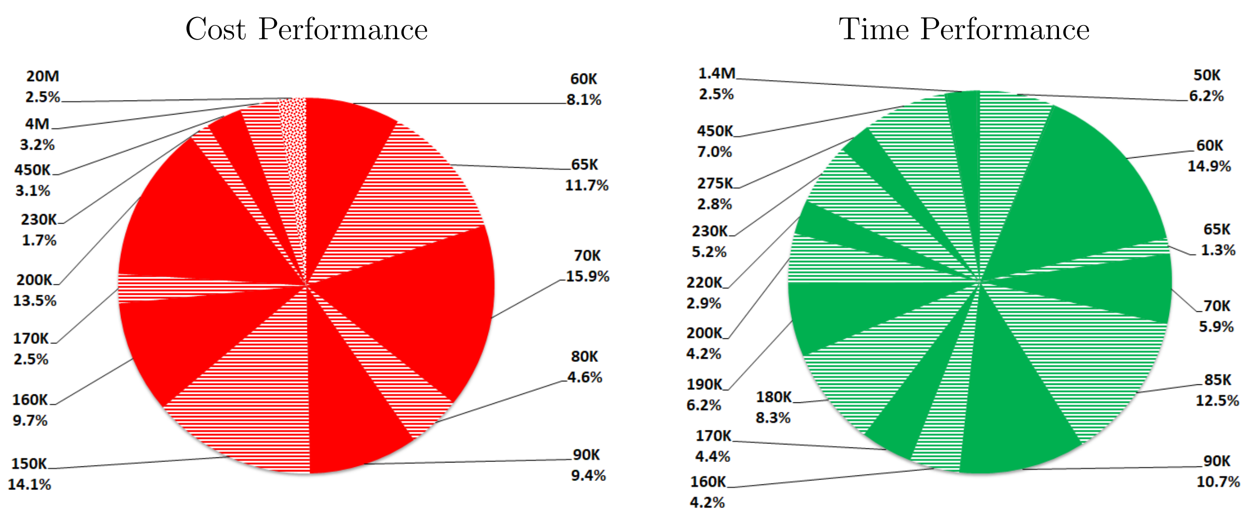
A.5 Synthetic Control Group Approach for Public Oversight

The SC method, pioneered by Abadie and Gardeazabal (2003) is an alternative approach for program evaluation that relaxes the parallel trends assumption. Abadie, Diamond and Hainmueller (2010) introduced the SC method in the context of comparative case studies, where only one or a few units are subject to intervention, while a larger set of units remained untreated. The central idea behind the SC method is that the outcomes from the control units are weighted so as to construct the counterfactual outcome for the treated unit, in the absence of the treatment. In contrast to the DD method, the SC method allows the effects of observed and unobserved predictors of the outcome to change over time, while assuming that pre-intervention covariates have a linear relationship with outcomes post-treatment. A SC unit is defined as the time-invariant weighted average of available control units, which prior to the intervention have similar pre-intervention characteristics and outcome trajectory to the treated unit. In order to run our SC robustness exercise, we thus need to i) specify the predictor variables employed for the SC group generation and ii) present the reasoning for defining the pool of controls.

First, when computationally feasible, perhaps the simplest strategy for specifying predictors is to include every pre-treatment outcome in the predictor set. Different sets of predictors may result in different synthetic controls, and there is little explicit guidance in the SC literature to assess predictor choice. Within the pre-intervention sample, one cannot do any better in terms of pre-intervention mean squared prediction error than to include every pre-intervention outcome. However, this will not be true when predicting out of the pre-intervention sample, which is ultimately the object of interest in SC design. Matching on high frequency - i.e., weekly or monthly - pre-intervention data may actually produce less reliable synthetic controls. As a result, as any characteristics unaffected by the policy treatment are possible candidates, we thus consider in this exercise annualized averages of the pre-treatment outcome (*Time Performance* and *Cost Performance*) and the control variables we use in our baseline DDD model (*Small, Negotiation, Fixed Price, Bureau Size*). Moreover, following the impact of bureaucratic competence on contract outcomes through red tape, as highlighted in section VI, we also include *time competence* and *cost competence* as predictors for *time performance* and *cost performance*, respectively.

Second, we split the lower control group in \$5,000 interval subgroups while the upper control group in larger intervals to include a similar number of observations for each: \$10,000 until \$250,000; \$25,000 between \$250,000 and \$500,000; \$50,000 between \$500,000 and \$1,000,000; \$100,000 between \$1,000,000 and \$1,500,000; \$250,000 between \$1,500,000 and \$3,000,000; \$500,000 between \$3,000,000 and \$4,500,000; \$1,000,000 between \$4,500,000 and \$10,000,000; \$5,000,000 between \$10,000,000 and \$30,000,000; those above \$30,000,000.

Figure A.7: Synthetic Control Group - Weights



Notes: Weights associated to each donor group in synthetic control group analysis for public oversight - cost (left panel) and time performance (right panel).

A.6 Propensity Score Matching DDD

In section V.3, we document a possible issue that can emerge from our analysis and undermine its validity. Indeed, the DDD method provides unbiased effect estimates under the assumption that in our data the construction and non-constructions trends over time would have been the same between the relative treatment and comparison groups in the absence of the intervention. By controlling for characteristics that may be contributing to unobservable selection bias, we are thus achieving identification of treatment effects. However, the treatment and control groups may also differ in ways that could affect their compositions overtime and DDD estimates may pick up differential shocks across subsectors or size of projects other than the ones we isolate (i.e., the fact that constructions are also subject to the Miller act) in addition to different trends. In order to account for this concern, we propose a test for robustness in which we combine the DDD methodology with a matching approach. By introducing a propensity score matching analysis we are appealing to the conditional independence assumption, that is the idea that matched comparisons imply balance on observed covariates, which recreates a situation similar to a randomized experiment where all subjects are essentially the same except for the treatment. In combination with the DDD strategy, we still account for of any unobserved characteristics that are constant across time between the two groups. The idea is to reconstruct the parametric DDD model presented in equation (3) as a weighted least squares procedure, where observations are weighted to ensure similarity on a set of observed pre-treatment variables. The DDD differences out the permanent confounders while matching on pre-treatment characteristics captures transitory shocks.[66] Specifically, we proceeded as follows. First, we compute the propensity score. The estimated propensity score $e(.)$ is the predicted probability of each contract of being treated and in our setting is derived from the fitted logit regression of the treatment status $Waiver_i$ on pre-treatment covariates $\mathbf{z}$.[67] We then construct a weighting variable w_i for the DDD es-

[66]Hybrid program evaluation methods like this are generally more robust than stand-alone approaches. See Cerulli (2016).

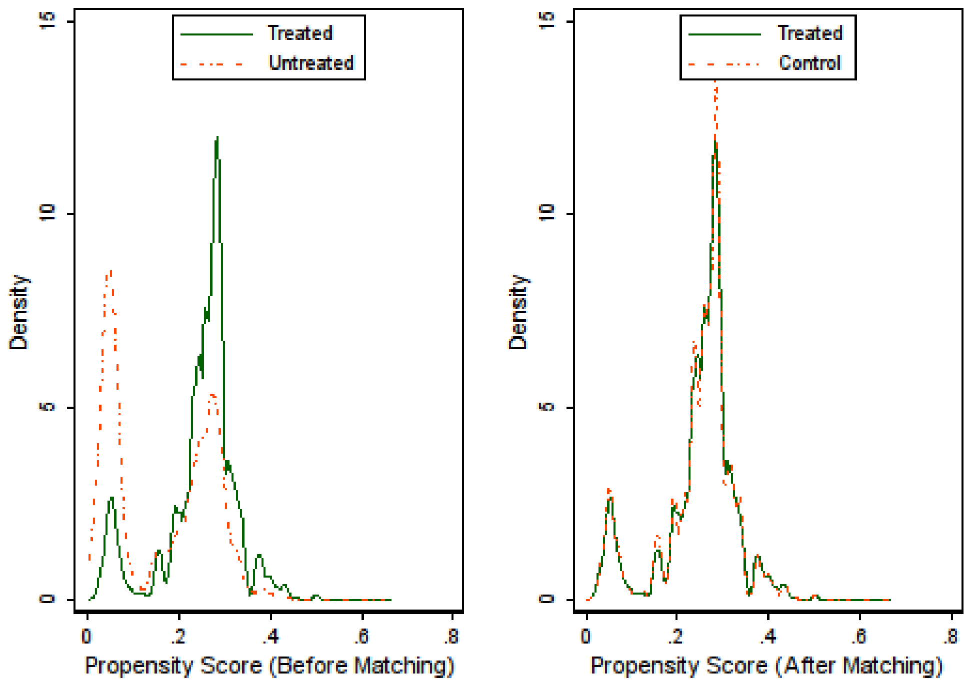
[67]We used as controls the 22 category fixed effects (2 for constructions and 20 for non-constructions), year fixed effects, *limited* and a dummy variable indicating whether the expected work duration exceeds the year length as a proxy for complexity. The choice mirrors the strategy to intercept intrasector time-varying shocks conditioning on different layers of object complexity. We also included the other exogenous controls from our outcome model and obtained comparable results.

timates similarly to an inverse probability of treatment weighting approach (DiNardo, Fortin and Lemieux, 1996). Specifically, following Hirano and Imbens (2001), the weights are calculated such that $w_i = Waiver_i + (1 - Waiver_i)\frac{e(\mathbf{z})}{1-e(\mathbf{z})}$ where the weight is equal to 1 for the treated group and $\frac{e(\mathbf{z})}{1-e(\mathbf{z})}$ for the untreated. The result of this operation was the creation of two groups of treated and untreated contracts with similar distribution of propensity score. The weights ensure that the covariates are uncorrelated with the treatment indicator. Each treated contract is matched with the the four closest untreated contracts based on similar values on the propensity scores.[68] In Figure (A.8), we show the comparison of the propensity score distributions across the treated and the untreated contracts before and after matching. This graph illustrates the successful adjustment of post-mastching propensity score distribution for the control group and visually shows that $e(\mathbf{z})$ in our analysis is in fact balanced across treatment and comparison groups.[69] We label this strategy as propensity score weighting triple difference. This estimator has the further advantage of not requiring the imposition of the linear-in-parameters form of the outcome equation. As such, it can be seen as a non-parametric DDD (see Heckman et al. (1998)).

[68] We follow Rosembaum and Rubin (1983) and use 4 matching neighbors and estimate the average treatment on the treated. We also perform the same analysis using 1, 2, 4, and 5 matching neighbors, finding similar results.

[69] We have also verified that covariates are balanced across treatment and comparison groups in the weighted sample. Table are available on request.

Figure A.8: Propensity Score Distribution: Pre-Post Matching Comparison - Cost Performance



Notes: Distributions of estimated probabilities of being a treated contract (propensity score). The comparison is presented by treatment group (blue line) and control group (red dashed line) prior to and after the matching is implemented.

A.7 Agencies and Clusters

Table (A.8) reports results on public and private oversight treatment effects for baseline DDD models of *Time Performance* and *Cost Performance* when standard errors are clustered at various levels: no clusterized robust (columns 1-2) - used in the empirical analysis -, *contractingoffice * year* (3-4), *contractingoffice * object* (columns 5-6), *contractingoffice * year * object* (columns 7-8), *contractingoffice * year * state* (columns 9-10) and *contractingoffice * year * state * object* (columns 11-12). DDD average treatment effects result in being robust to these different clusterization levels for both public and private oversight.

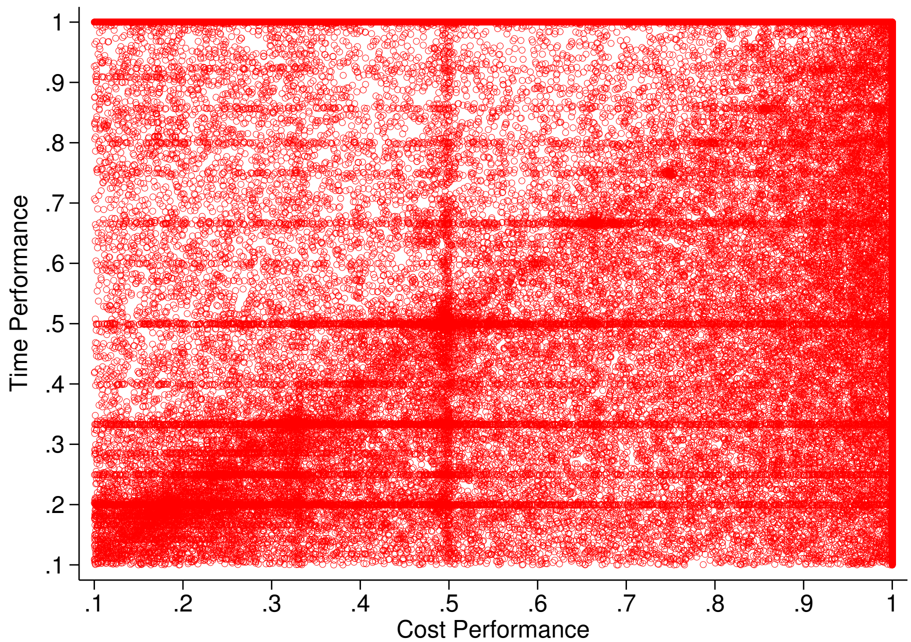
Table A.8: Baseline model with different cluster level

	robust		off*year		off*obj		off*year*obj		off*year*state		off*year*state*obj	
	T	C	T	C	T	C	T	C	T	C	T	C
$\hat{\theta}_{public}$	0.0476	0.0430	0.0476	0.0430	0.0476	0.0430	0.0476	0.0430	0.0476	0.0430	0.0476	0.0430
	(0.00600)	(0.00493)	(0.0148)	(0.0143)	(0.0188)	(0.0157)	(0.0129)	(0.0107)	(0.0104)	(0.00891)	(0.00957)	(0.00808)
$\hat{\theta}_{private}$	-0.0705	-0.0392	-0.0705	-0.0392	-0.0705	-0.0392	-0.0705	-0.0392	-0.0705	-0.0392	-0.0705	-0.0392
	(0.0257)	(0.0142)	(0.0279)	(0.0217)	(0.0395)	(0.0225)	(0.0337)	(0.0191)	(0.0286)	(0.0175)	(0.0283)	(0.0169)
N	99,578	100,861	99,578	100,861	99,578	100,861	99,578	100,861	99,578	100,861	99,578	100,861
# Clusters	#	#	1052	1052	1466	1466	8511	8511	14560	14560	43098	43098

Notes: results of various DDD models of *Time Performance* - odd columns - and *Cost Performance* - even columns - on public and private oversight treatment indicators plus controls for firm size, tender type, whether the firm is a limited liability company, whether the contract was signed during the last week of the fiscal year, bureau, year, state, and contract object fixed effects. Standard errors are clustered at various levels: no clusters (columns 1-2), *contractingoffice * year* (3-4), *contractingoffice * object* (5-6), *contractingoffice * year * object* (7-8), *contractingoffice * year * state* (9-10) and *contractingoffice * year * state * object* (11-12). For each clusterization level, *# Clusters* report the relative number of clusters in the data.

A.8 Additional Tables and Figures

Figure A.9: Scatterplot of Performances Measures



Notes: Each point represents the contract level of *Time* and *Cost Performance*.

Table A.9: Triple Difference - Time Performance

Panel (a): Time Performance

	(1)	(2)	(3)	(4)	(5)
$\hat{\theta}_{public}$	0.026	0.028	0.050	0.050	0.048
	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)
$\hat{\theta}_{private}$	-0.001	0.004	-0.064	-0.066	-0.067
	(0.025)	(0.025)	(0.024)	(0.026)	(0.027)
Waiver	0.092	0.076	0.053	0.052	0.050
	(0.008)	(0.007)	(0.006)	(0.006)	(0.005)
Constr	-0.014	-0.031	0.002	0.000	0.017
	(0.018)	(0.018)	(0.011)	(0.011)	(0.007)
Constr * Waiver	-0.017	-0.007	0.020	0.023	0.025
	(0.019)	(0.017)	(0.016)	(0.016)	(0.015)
Constr * Post	0.022	0.023	0.038	0.037	0.036
	(0.023)	(0.022)	(0.015)	(0.015)	(0.013)
Small		-0.011	-0.004	-0.004	-0.004
		(0.006)	(0.005)	(0.005)	(0.005)
Negotiation		0.008	-0.015	-0.017	-0.019
		(0.011)	(0.007)	(0.007)	(0.006)
Fixed Price		0.085	0.070	0.070	0.076
		(0.011)	(0.009)	(0.008)	(0.008)
Bureau Size		-0.000	-0.000	-0.000	-0.000
		(0.000)	(0.000)	(0.000)	(0.000)
N	98,089	98,089	98,089	98,089	98,089
R^2	0.015	0.028	0.094	0.098	0.110
$Avg_{services}$	0.763	0.763	0.763	0.763	0.763
Avg_{works}	0.763	0.763	0.763	0.763	0.763
Controls		✓	✓	✓	✓
Bureau Fixed Effects			✓	✓	✓
State Fixed Effects				✓	✓
Object Fixed Effects					✓

Notes: Notes: complete set of regressions coefficients of Table (3) panel a).

Table A.10: Triple Difference - Cost Performance

	Panel (b): Cost Performance				
	(1)	(2)	(3)	(4)	(5)
$\hat{\theta}_{public}$	0.018	0.016	0.039	0.039	0.045
	(0.009)	(0.010)	(0.011)	(0.011)	(0.010)
$\hat{\theta}_{private}$	-0.003	0.004	-0.028	-0.030	-0.037
	(0.013)	(0.014)	(0.016)	(0.016)	(0.015)
Waiver	0.102	0.065	0.043	0.042	0.040
	(0.007)	(0.008)	(0.006)	(0.006)	(0.006)
Constr	0.162	0.099	0.114	0.106	0.072
	(0.022)	(0.019)	(0.014)	(0.013)	(0.008)
Constr * Waiver	-0.057	-0.020	-0.007	-0.005	-0.001
	(0.010)	(0.010)	(0.009)	(0.008)	(0.008)
Constr * Post	-0.045	-0.029	-0.012	-0.012	-0.013
	(0.026)	(0.020)	(0.016)	(0.015)	(0.012)
Small		-0.005	0.008	0.008	0.009
		(0.005)	(0.004)	(0.004)	(0.004)
Negotiation		0.009	-0.004	-0.007	-0.002
		(0.008)	(0.006)	(0.006)	(0.005)
Fixed Price		0.223	0.185	0.184	0.174
		(0.010)	(0.009)	(0.009)	(0.009)
Bureau Size		-0.000	0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)
N	98,089	98,089	98,089	98,089	98,089
R^2	0.036	0.120	0.184	0.189	0.206
$Avg_{services}$	0.832	0.832	0.832	0.832	0.832
Avg_{works}	0.918	0.918	0.918	0.918	0.918
Controls		✓	✓	✓	✓
Bureau Fixed Effects			✓	✓	✓
State Fixed Effects				✓	✓
Object Fixed Effects					✓

Notes: complete set of regressions coefficients of Table (3) panel b).