Online supplement - not for publication

This online supplement consists of five separate appendices (Appendix A - E). Each Appendix provides auxiliary material for one of the main sections of our article, Hyytinen, Lundberg and Toivanen "Design of public procurement auctions: Evidence from cleaning contracts".

Appendix A: Auxiliary material to Section 2

This Appendix has two parts. In Part A we provide additional information on the regulation of the choice of the supplier in the Swedish procurement auctions. In Part B we offer a summary of the mandatory qualification criteria and scoring rules used in the new regime.

Part A: Description of procurement legislation

Regulation during the old regime

The Public Procurement Act ("Lag 1992:1528 om offentlig upphandling, LOU") was not yet in force in 1990-1993, but the rules that applied then were essentially the same as under the act. The law was based on the EU rules that prevailed at the time: The Swedish law followed the COUNCIL DIRECTIVE 92/50/EEC of 18 June 1992 relating to the coordination of procedures for the award of public service contract.

What follows is an excerpt from COUNCIL DIRECTIVE 92/50/EEC of 18 June 1992 relating to the coordination of procedures for the award of public service contract: CHAPTER 3, Criteria for the award of contracts, Article 36:

1. *Without prejudice to national laws, regulations or administrative provisions on the remuneration of certain services, the criteria on which the contracting authority shall base the award of contracts may be:*

a. *where the award is made to the economically most advantageous tender, various criteria relating to the contract: for example, quality, technical merit, aesthetic and functional characteristics, technical assistance and after-sales service, delivery date, delivery period or period of completion, price; or*

b. *the lowest price only.*

2. *Where the contract is to be awarded to the economically most advantageous tender, the contracting authority shall state in the contract documents or in the tender notice the award criteria which it intends to apply, where possible in descending order of importance.*[1]

Regulation during the new regime

The new Public Procurement Act ("Lag 2007:1091 om offentlig upphandling, LOU") became effective as of January 1, 2008. It was a direct consequence of the 2004 EU procurement directive, 2004/18/EC (Directive 2004/18/EC of the European Parliament and of the Council of 31 March 2004 on the coordination of procedures for the award of public works contracts, public supply contracts and public service contracts).

What follows is an excerpt from Directive 2004/18/EC of the European Parliament and of the Council of 31 March 2004 on the coordination of procedures for the award of public works contracts, public supply contracts and public service contracts, Section 3, Award of the contract, Article 53, Contract award criteria:

1. *Without prejudice to national laws, regulations or administrative provisions concerning the remuneration of certain services, the criteria on which the contracting authorities shall base the award of public contracts shall be either:*

[1] This clause was effectively neither followed nor enforced, enabling the municipalities to run beauty contests.

a. *when the award is made to the tender most economically advantageous from the point of view of the contracting authority, various criteria linked to the subject-matter of the public contract in question, for example, quality, price, technical merit, aesthetic and functional characteristics, environmental characteristics, running costs, cost-effectiveness, after-sales service and technical assistance, delivery date and delivery period or period of completion, or*

b. *the lowest price only.*

2. *Without prejudice to the provisions of the third subparagraph, in the case referred to in paragraph 1(a) the contracting authority shall specify in the contract notice or in the contract documents or, in the case of a competitive dialogue, in the descriptive document, the relative weighting which it gives to each of the criteria chosen to determine the most economically advantageous tender.*

Those weightings can be expressed by providing for a range with an appropriate maximum spread. Where, in the opinion of the contracting authority, weighting is not possible for demonstrable reasons, the contracting authority shall indicate in the contract notice or contract documents or, in the case of a competitive dialogue, in the descriptive document, the criteria in descending order of importance.

Part B: Description of entry modes, the mandatory qualification and exclusion criteria and scoring rules

Entry modes

Both the old and new procurement laws allowed for four types of entry modes, called Simplified, Open, Restricted, Negotiated. The main difference between these four modes is that Simplified and Open allowed free entry, although the other two (Restricted, Negotiated) did

not. Although negotiations were allowed in some types of procurements, they were not used in the procurements that we study.

Our indicator for free entry takes value of one when the entry mode of an auction was Simplified or Open, and is zero otherwise (i.e., when entry was restricted).

Mandatory qualification and exclusion criteria and scoring rules

In the old regime, there were no formal mandatory qualification and exclusion criteria or explicitly announced scoring rules. In the new regime, the law allowed only first price or scoring auctions, and formal mandatory qualification and exclusion criteria were used widely.

We do not have comparable data on the scoring rules for all the scoring auctions of the new regime. There are three reasons for this: i) For some scoring auctions the scoring rule information was partly missing or the rule was incompletely defined; ii) the scoring rules can take different forms and are not entirely comparable across the scoring auctions (see also below); and iii) we cannot always reliably distinguish whether a particular clause was related to the scoring rule (award criterion) or whether it was one of the mandatory qualification and exclusion criteria.

We have analyzed the available (somewhat scattered) data on mandatory qualification and exclusion criteria and scoring rules as follows: Building on Lundberg et al. (2015), we first grouped the criteria into two main categories: Non-environmental quality criteria ("Quality criteria") and Environmental criteria ("Environmental criteria"). In the original data there were 26 different kinds of quality criteria and 28 environmental quality criteria. On average, the cleaning service auctions of the new regime used about ten quality criteria and six environmental criteria. Many of these were related to the mandatory qualification and exclusion criteria (see below), but for a number of cases we cannot say exactly which. Moreover, some of the criteria were very similar and had similar kind of content. Therefore, following Lundberg

et al. (2015), we aggregated both the various environmental criteria and the various quality criteria into six main variables, respectively. The six quality variables are “Financial status, FIN” (e.g., data on supplier’s financial condition); “Insurance, INS (e.g. documentation of the supplier having required insurance); “Experience, EXP (e.g., documentation of relevant experience in similar assignments); “Performance plan, PER” (e.g., a description of how the cleaning service contract will be carried out); “Social criteria, SOC” (e.g. documentation of supplier having collective labor agreements with the union, etc); and ”Staffing, STAF” (e.g., documentation describing the qualifications of the employees). The six environmental variables are “Environmental management system, EMS” (e.g., having environmental certificates and/or fulfilling different ISO 14000 standards), “Eco labeling ECO” (e.g., having/using the EU Ecolabel for the cleaning products); “Vehicle, VEH” (e.g. meeting of certain emission standards for cars); “Meeting various chemical-usage and environmental codes, CHEM” (e.g., meeting the Swedish Chemicals Agency Code of Statutes 2008 and similar regulations); “Eco monitoring, MON” (Intention of the authority to monitor that the supplier meets the required environmental standards); and “Other eco demands, OTHER “ (other types of environmental or allergy-related criteria).

Table A1.1 provides summary statistics for the quality and environmental variables. It shows that except for SOC, most of the quality criteria were used in more than 90% of the auctions (incl. first price auctions) of the new regime. This finding means that many of them were mandatory qualification and exclusion criteria.

Table A1.1: Quality criteria

Variable	Obs	Mean	Std. Dev.	Min	Max
Panel A: Quality criteria					
FIN	355	0.98	0.14	0	1
INS	355	0.95	0.21	0	1
EXP	355	0.97	0.18	0	1
PER	355	0.99	0.07	0	1
SOC	355	0.54	0.50	0	1
STAF	355	0.94	0.25	0	1

Panel B: Environmental criteria					
MON	355	0.14	0.34	0	1
ECO	355	0.79	0.40	0	1
EMS	355	0.69	0.46	0	1
VEH	355	0.25	0.43	0	1
CHEM	355	0.87	0.34	0	1
OTHER	355	0.55	0.50	0	1

Notes: This table summarizes quality and environmental criteria used either as the mandatory qualification and exclusion criteria or as a part of the scoring rule. The sample includes both scoring auctions and first price auctions. The exact data on the scoring rule formulae are not available for all the scoring auctions.

A key finding from the above description is that due to the level of detail of the criteria imposed by the municipalities, there was effectively very little room for non-contractible quality. It also seems that many of the listed criteria were mandatory qualification and exclusion criteria, and not parts of the scoring rules. Moreover, compared to more complex products (e.g. construction projects), these procurement criteria are relatively simple. Many of them were related to the supplier meeting a certain environmental standard or to (easy-to-verify) aspects of its business operations, such as its financial condition or basic employee qualifications.

Bergman and Lundberg (2013) provide a detailed analysis of the various types of scoring rules used in Sweden. We briefly discuss them in Appendix D.

References used in this Appendix:

Bergman, M. A. and Lundberg, S. “Tender evaluation and supplier selection methods in public procurement.” Journal of Purchasing and Supply Management, Vol.19 (2013), pp. 73-83.

Lundberg, S., Marklund, P-O. Strömbäck, E. and Sundström, D. "Using public procurement to implement environmental policy: an empirical analysis." Environmental Economics and Policy Studies, Vol. 17 (2015), pp. 487-520.

Appendix B: Auxiliary material to Section 4

In this Appendix, we report a set of reduced-form regressions to which we refer in Section 4 of the main article. We use these reduced-form regressions to explore how the regime change is associated with the winning bids.

Table A2.1 shows how procurement costs vary between the two regimes. It displays regressions in which the dependent variable is the logarithm of the winning bid and which include the regime dummy (R = 1 if regime is new; and = 0 otherwise), or auction format dummies, as the key regressor(s). The table reports results for six different model specifications. Model 1 includes only the regime dummy. In Model 2 we add the two groups of explanatory variables. The first group of variables, W, includes the number of contracts (*Auctions per procurement*) awarded in the procurement event, the size (in 10 000 square meters) of the premises covered by the contract (*Size of premises*) and its square, the length (in years) of the contract (*Contract length*), the number of years over which the contract can optionally be extended if the municipality so decides when the initial contracting period expires (*Extension*), and the scaled number of days during which the cleaning takes place (*Frequency*) and its square. The second group of control variables, X, consists of dummies for types of premises covered by a contract (four categories: *School*, *Office*, *Day-care center*, *Other*), municipal unemployment rate (*Unemployment*), population density (*Population density*) in thousands of inhabitants per the geographical size of the municipality, share of inhabitants having a higher education (*Education*) and a binary indicator that obtains a value of one for those municipalities where leftwing parties have more than 50 percent of the seats in the municipal council (*Red majority*). We also include a linear within-sample trend (T). Model 3 is equivalent to Model 2, except that it also includes municipal fixed effects.

As the table shows, the winning bids do not differ across the two regimes in Models 1-3. The coefficient of the regime dummy is positive, but has a large standard error and is thus not significant.

Table A2.1: Reduced form regressions, dependent variable = Ln(winning bid), OLS

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Regime-dummy, R	0.043	0.111	0.105	-	-	-
	(0.063)	(0.172)	(0.159)			
Scoring	-	-	-	0.052	0.111	0.114
				(0.074)	(0.170)	(0.158)
First price	-	-	-	0.026	0.112	0.051
				(0.069)	(0.188)	(0.179)
Control variables (W, X, T)	No	Yes	Yes	No	Yes	Yes
Municipal FE	No	No	Yes	No	No	Yes
R2	0.002	0.321	0.479	0.002	0.321	0.479
Observations	1,075	1,067	1,067	1,075	1,067	1,067

Notes: The standard errors reported in parentheses are clustered by procurements. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

We also report in the table results from Models 4 to 6: These models are equivalent to Models 1 to 3, except that they do not force the coefficient of *Scoring* auction be equal to the coefficient of *First price* auction. When we substitute the indicators of the two auction formats for the regime indicator, the null hypothesis that the procurement costs were equal across the three auction formats cannot be rejected in any of Models 4 to 6. For example, in Model 6, p-value of the test statistic is 0.52. We cannot reject the null hypothesis that the coefficients of the *First price* and *Scoring* -dummies are equal either. For example, the p-value of the test statistic is 0.40 in Model 6.

In sum, the reduced-form models reported here provide no evidence that the winning bids were different across the old and new regimes. We have re-run these reduced form-models also in two alternative ways: First, we included our instruments (see main text) as additional regressors. Second, instead of controlling for a linear within-sample time trend, we either dropped the trend variable altogether or added its square to the models. These alternative regressions confirm that procurement costs were not lower in the new regime.

Appendix C: Auxiliary material to Section 5

In this Appendix we report a number of robustness tests for IV estimations of equations (1) and (2), reported in Section 5 of the main text. The Appendix consists of three parts: Part A: explores in detail whether our instruments are weak. Part B studies winner type and entry effect (composition). Finally, in Part C we report details of the other robustness tests that we have implemented.

Part A: Weak instruments

As we reported in Table 2 of the main body of the article, the F-test for the joint significance of our instruments is 13.56 in our preferred IV model (Model 1 of Table 3). This exceeds ten, which is the threshold often used to detect weak instruments (Staiger and Stock, 1997; Stock, Wright and Yogo, 2002). This finding supports the view that, when considered as a bundle, there is variation in the instruments that allows predicting the endogenous variable, even after conditioning on the other observables and the municipal fixed effects.

To err on the conservative side, we have explored further the possibility that our instruments are weak. We have implemented the following analyses:

First, we applied Moreira's (2003) conditional likelihood-ratio test (CLR-test) to our preferred IV model. The CLR-test allows us to test the null hypothesis that the coefficient of the number of submitted bids is zero in the second stage, *without assuming that the instruments are strong*. The CLR-test rejects the null hypothesis firmly (p-value < 0.01). The weak-instrument-robust 95% confidence intervals associated with the statistic is [-1.09, -0.17]. Roughly put, this is the set of parameter values for α that are consistent with the data when we allow the instruments to be weak. More formally, the 95% confidence interval refers to those coefficient values for which the rejection probability is below 95%. We conclude that the

documented negative effect of actual entry on the winning bid is not an artifact of weak instruments.

Second, we have checked the partial R2 statistic in the first-stage of our IV (see, e.g., Bound, Jaeger, and Baker 1995). In our set up, the statistic measures the correlation between the logarithm of actual entry and the instruments after partialling out the effect of the included exogenous variables. The benefit of using the partial R2 statistic is that it accounts for the possibility that the included exogenous variables predict actual entry strongly. Standard R2 or adjusted R2 would in such a case give a too promising picture of the capability of the first-stage to explain variation in the endogenous variable. In our preferred IV specification, the partial R2 statistic is equal to 0.114. This compares quite favorably to the adjusted R2 from the first-stage, which suggests that overall, we explain about 64% of the variation of the logarithm of actual entry. We point out that a large fraction of the explanatory power in the first-stage comes from the municipal fixed effects: Were the municipal fixed effects not included in the first-stage, the partial R2 statistic would be equal to 0.20 and the adjusted R2 about 40%. To our best knowledge, there are no clear-cut criteria in the literature on how large the partial R2 statistic ought to be, but our number appears not to be particularly low.

Third, a standard, yet somewhat informal way to check how IV works is to explore the reduced form of IV, i.e., the regression of the ultimate outcome variable (here: ln(winning bid)) on the instruments and the controls. We found that the three instruments are jointly highly significant (F = 7.19, p-value of the F-test = 0.0001) in this reduced-from regression. The instruments also obtained the expected signs: Free entry obtains a negative coefficient, as does the number of potential bidders. These patterns are consistent with our IV results.

Fourth, we have checked our (over-identified) 2SLS estimates against Limited information maximum likelihood (LIML) -estimates, using our preferred baseline model from Table 3 (i.e. Model 1). The motivation to do this is that the LIML-estimator has better small

sample properties than 2SLS when instruments are weak (see, e.g., Stock, Wright, and Yogo 2002). The LIML estimates of the effect of entry on the winning bids look very similar and the standard errors are not larger (coeff. = -0.60, p-value < 0.001). These findings suggest that our IV findings are not driven by a small sample bias, if the instruments happened to be weak.

Sixth, the biases of IV-estimators may increase with the number of instruments (Hahn and Hausman 2003). It has been argued that just-identified LIML IV is approximately median-unbiased even with weak instruments. Moreover, in the just-identified case, LIML and 2SLS are equivalent. Therefore, running the IV using the best single instrument is yet another robustness test that we can implement. When we use the Free entry-indicator (but not the interaction, nor Ln(N)) as the instrument, the free entry is highly significant and positive in the first stage of the IV (coeff = 0.41, p-value < 0.0001). Moreover, the coefficient of the submitted number of bids in the second stage is in line with our baseline findings, and in fact more negative (coeff = -0.84, p-value = 0.0001). This suggests that if anything, our baseline IV findings may be biased towards zero and thus conservative.

Finally, as Stock and Yogo (2005) explain, an alternative consequence of weak instruments (i.e., poor predictive ability of the instruments) is that when a parameter is estimated by IV-estimators, hypothesis tests concerning the estimated parameter suffer from size distortions. Using this as an insight, they develop an explicit test of weak instruments: Instruments are weak if a Wald test at the 5% level can have an actual rejection rate of no more than a given threshold (e.g., 10%, or 15%). To implement the test, suppose that we test the null hypothesis that actual entry has no effect on the winning bids and that we accept a rejection rate of 10% (which is inflated relative to the nominal 5% rate of the Wald test). The tabulations in Stock and Yogo (2005) show that we can reject the null hypothesis of weak instruments if a minimum eigenvalue statistic obtains a value larger than 22.30. In our case, the statistic is 43.20. We thus reject the null hypothesis of our instruments being weak also using this approach.

To conclude, instruments are weak if, conditional on the included regressors, they don't have sufficient variation to explain variation in the endogenous variable. The analyses reported above show that such lack of variation is not a problem in our empirical set up.

Part B: Winner type and entry effect (composition)

Could the negative effect of actual entry on bids be driven by a compositional change among the winners? To explore this, we have re-run Models 1 to 3 of Table 3 of the main text after including supplier type dummies as additional controls. Following Marion (2009), these estimations are meant to uncover how much of the entry effect is due to within firm-type effects (i.e., lowering the bids of particular supplier type) and how much is due to a compositional effect (i.e., changing which type of supplier wins the auction). These estimations (see Table A3.1) show that the elasticity estimate does not change when supplier type dummies are included. This finding provides support for the view that the estimated entry effect is due to additional entry inducing more aggressive bidding by suppliers and is *not* due to different *types* of suppliers winning in auctions with more entry.

Table A3.1: Log(winning bid), firm type effects included

	2SLS		
	Model 1	Model 2	Model 3
Ln(n)	-0.515***	-0.558***	-0.558***
	(0.134)	(0.140)	(0.140)
Scoring	-0.083	-0.070	-0.070
	(0.180)	(0.177)	(0.177)
First price	-0.083	-0.298	-0.297
	(0.180)	(0.223)	(0.223)
Control variables (W, X)	Yes	Yes	Yes
Municipal FE	Yes	Yes	Yes
Firm type dummies	Yes	Yes	Yes
Observations	1,067	1,067	1,067

Notes: Cluster-robust standard errors in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.10. The baseline estimations for these regressions are those reported in Table 3 of the main text. Variable Ln(n) refers to actual entry, i.e., the logarithm of the number of submitted bids.

Part C: Other robustness tests

We have collected the results from the rest of our robustness analyses (#1-#5) to Table A3.2. Taking each of them in turn:

Exclusion restriction (robustness test #1): We evaluate the *plausibility* of our exclusion restrictions using the approach developed by Conley et al. (2012). This approach can be applied if one suspects that the exclusion restriction of the instruments does not hold *exactly*. In our case, this would amount to arguing that one or more of our instruments F = {*Free entry*, *Free entry* × *R*, *Ln*(*N*)} are correlated with the unobservables influencing the winning bid.

To explain the method briefly, let γ be a (3 x 1) vector of parameters that measures whether the instruments, F, can be excluded from the second stage of 2SLS. If $\gamma = (0, 0, 0)$ exactly, the variables have no direct effect on the winning bids. If we allow a 'prior' distribution for γ, we can discuss how close the exclusion restriction is to being satisfied. If it is likely that γ is close to zero and if the probability that it is far away from zero decreases sufficiently rapidly, we can say that the instruments are plausibly exogenous. We can then explore how robust the results are to small violations of the exclusion restriction. We have implemented the local-to-zero method of Conley, Hansen and Rossi (2012), as follows. First, our prior is that γ is normally distributed. We allow for the following means of the prior distribution: {(0.1, 0.1, 0.1), (-0.1, -0.1, -0.1), (0.1, 0.1, 0.0), (0.0, 0.0, -0.1)}. For example, (0.1, 0.1, 0.0) is equivalent to assuming that the use of free entry was positively correlated with the unobservable determinants of the procurement costs, but potential entry was not. The variance of γ is set to $(0.1^2, 0.1^2, 0.1^2)$.

Table A3.2 displays the results for various implementations of the local-to-zero method of Conley et al. (2012). It shows that the coefficient of actual entry varies a little, stays in the range [-0.6, -0.5] and remains significant at better than 5% level in each case. Our IV results are hence robust to correlation of our instruments with the unobservables influencing the

winning bids. For example, robustness test #1-c shows that our results hold even if municipalities had been more likely to allow free entry into auctions in which the expected procurement costs were higher.

We have also checked that our results are robust to the possibility that restricted entry was used in the old regime to deter unreliable bidders who were expected to bid *too low* (see Decaloris 2014). This amounts to checking whether the use of free entry in the old regime was positively associated with the unobserved determinants of the winning bids. We therefore set the means of the prior distribution of γ to (0.2, 0.0, 0.0). When we do so, we find that the coefficient of $Ln(n)$ is -0.57 (p-value = 0.011). Our IV results are robust to the possibility that restricted entry was used to deter unreliable bidders in the old regime.

Table A3.2: Robustness checks

	Coefficient of Ln(n) (i.e., a)	
#1: Exclusion restriction (Conley et al. 2012)		
a: Local deviation: γ = (0.1, 0.1, 0.1)	-0,569	**
	(0,224)	
b: Local deviation: γ = (-0.1, -0.1, -0.1)	-0,521	**
	(0,224)	
c: Local deviation: γ = (0.1, 0.1, 0.0)	-0,559	**
	(0,224)	
d: Local deviation: γ = (0.0, 0.0, -0.1)	-0,534	**
	(0,224)	
#2: Alternative measure of potential entry	-0,696	***
	(0.193)	
#3: Alternative estimation sample (623 auctions)	-0,579	***
	(0.292)	
#4: Outliers (winzorized at +/- 3 std)	-0,331	***
	(0.105)	
Municipal FE	Yes	
Control variables	Yes	

Notes: The standard errors reported in parentheses are clustered by procurements. *** p < 0.01, ** p < 0.05, * p < 0.10. The baseline for these robustness checks is Model 1, reported in Table 3 of the main text.

Alternative measure of potential entry (robustness test #2): We have in our baseline analysis measured potential entry using the total number of suppliers that submitted at least one bid in a given municipality in our data, calculated separately for large and small contracts. The threshold for a large contract has so far been the 80th percentile of the distribution of the total size of the premises, calculated separately for each municipality. If we use the median instead, we find that in the first stage, the coefficient of potential entry decreases a little but remains significant. The IV estimate from the second stage, in contrast, becomes a bit more negative (increases in absolute value) and remains significantly different from zero.

Alternative estimation sample (robustness test #3): In our baseline estimations we include municipal fixed effects for the municipalities that organize procurements in the old or in the new regime. If we restrict our estimation sample to only those municipalities that organized procurements during both periods, the size of our estimation sample decreases considerably, to 623 auctions. When we repeat the 2SLS analyses reported in Table 3, we find that our results are robust to using this smaller sample. For example, as Table A3.2 shows, the entry elasticity of the winning bids is -0.58 for our preferred IV model. Despite the smaller sample, the estimate is still significant at better than 5% level.[2]

Outliers (robustness test #4)*:* We have also checked that very large or small winning bids are not driving our findings. There are many ways to identify outliers and to explore their effects, but we identified them as those bids that were larger (or smaller) than three standard deviations from the sample mean. We then dropped those observations from the estimation sample. The results from estimating our preferred IV model using this trimmed sample are

[2] We would like to point out two things about the smaller sub-sample. First, the municipalities which remain in this sub-sample organized price only auctions in the new regime less often than in the full sample. In the full sample, the share of price only auctions is about one third. In the sub-sample of the municipalities that organized procurements during both periods of our data, the share is about one fifth. Second, potential entry no longer predicts actual entry in this (much) smaller sample. However, *Free entry* works as in the full sample.

reported in the table above: As can be seen, our IV estimate decreases, but remains negative and significant. In an alternative analysis, we kept the large and small observations in the sample, but set them to be equal to the value they would obtain had they been exactly three standard deviations from the sample mean. In this (unreported) estimation, the IV estimate of the coefficient of $Ln(n)$ is -0.48 (p-value < 0.001).

Different trend specification: Finally, in our baseline analysis, we control for a within-sample (linear) time trend. Our IV estimations (reported in Tables 2 and 3 of the main text) are also robust to dropping the trend variable altogether; and to adding its square to the model.

References used in this Appendix:

Bound, J., Jaeger, D. A. and Baker, R. M. "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak". Journal of the American Statistical Association, Vol. 90 (1995), pp. 443–450.

Conley, T. G., Hansen, C. B. and Rossi, P. E. "Plausibly exogenous." Review of Economics and Statistics, Vol. 94 (2012), pp 260-272.

Decarolis, F. 2014. "Awarding price, contract performance, and bids screening: Evidence from procurement auctions." American Economic Journal: Applied Economics, Vol. 6 (2014), pp. 108-132.

Hahn, J. and Hausman, J. A. "Weak instruments: Diagnosis and cures in empirical econometrics". American Economic Review Papers and Proceedings, Vol. 93 (2003), pp. 118–125.

Li, T. and Zheng, X. "Entry and competition effects in first-price auctions: Theory and evidence from procurement auctions." Review of Economic Studies, Vol. 76 (2009), pp. 1397-1429.

Marion, J. "Are bid preferences benign? The effect of small business subsidies in highway procurement auctions." Journal of Public Economics, Vol. 91 (2007), pp. 1591-1624

Moreira, M. J. "A conditional likelihood ratio test." Econometrica, Vol. 71 (2003), pp. 1027-1048.

Staiger, D. and Stock, J. H. “Instrumental Variables Regression with Weak Instruments,” Econometrica, Vol. 65 (1997), pp. 557 – 586.

Stock, J. H. and Yogo, M. “Testing for weak instruments in linear IV regression”. In Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg, ed. D. W. K. Andrews and J. H. Stock, New York: Cambridge University Press, pp. 80–108, 2005.

Stock, J. H., Wright, J. H. and Yogo, M. “A survey of weak instruments and weak identification in generalized method of moments” Journal of Business and Economic Statistics Vol. 20 (2002), pp. 518–529.

Appendix D: Auxiliary material to Section 6

This Appendix consists of five parts: In Part A, we provide a brief summary of how we calculated the price weights. In Part B, we provide a summary table of the interview questions to which we refer in Section 6 of the main text. In Part C, we describe the random utility model (Conditional logit model), which we use in the main text to analyze the price sensitivity of the municipalities. In Part D we report further analyses using quantile regressions with the aim to explore 1) whether the bid distributions were different in the two regimes and 2) how other bidders reacted to the presence of an *inhouse* bidder. Finally, in Part E we detail our investigation based on Bandiera, Pratt and Valletti (2009) on passive versus active waste.

Part A: Calculation of price weights

We have also analyzed the exact formulae of the scoring rules.[3] These data were obtained from the calls of tenders. A scoring rule can require a scaling of the price (bid) to match the quality measure or, alternatively, transforming of the quality measure to monetary units (scaling to price).[4] These rules have in practice different designs and mathematical formulae. For example, in our data a commonly used scoring rule was the one in which the "quality-equivalent of the monetary bid" was obtained by multiplying the maximum obtainable quality score by the ratio of the lowest submitted bid to bidder *i*'s bid (i.e., $S_i = S_{max} \times (Bid_{Lowest} / Bid_i)$). This score measures the monetary value of supplier *i*'s bid in units of the quality measure. This scaled bid and the actual quality score of supplier *i* were then given weights (summing to one) when the final total score was calculated. This was used in 40.3% of the scoring auctions in our data.

[3] Our article's data for the new regime are a subset of the data analyzed by Bergman and Lundberg (2013), who explore the various forms and designs of scoring rules in more detail. Our analysis here builds on this prior work.

[4] A third type of scoring auction is "quality only scoring". In this case suppliers compete only in the quality dimension and the contracting authority announces a fixed payment in the call for tender. We have only one such example in our data.

The weight that a scoring rule attached to the price is not available for all auctions. Nor are the weight data entirely comparable across the auctions, because the scoring rules differ between auctions. We have comparable weight data for about 73% of the scoring auctions of the new regime. Using these data, we find that the average weight in the scoring auctions was 0.49. In an auction that chooses the winner solely based on price, the weight on the monetary bid is (normalized) to one. Taking into account missing data on the scoring rules, the average weight attached to the price was then $0.49\times(2/3)\times0.73 + 1\times(1/3)\times(1/0.73) \approx 0.70$. If we assume that the average weight in the scoring auctions for which the weights are not observed was the same as in the other scoring auctions, the weight attached to the price in the new regime would be $0.49\times(2/3) + 1\times(1/3) \approx 0.66$.

Part B: Summary of the interview data

In this part of the Appendix, we provide a summary table of the auxiliary interview questions used in Section 6.

Table A4.1 Interview evidence (referred to in Section 6)

Q1	Reasons for not using formal scoring rules in the old regime (share of respondents indicating the reason, many choices allowed)	
	Didn't have experience to use formal scoring rules	0.91
	Convenience	0.86
	Lobbying by bidders	0.18
	Decision by municipal politicians	0.18
	It gave the freedom to pick the winner	0.77
	Other, please explain	0.09
Q2	Reasons for not choosing the lowest bid in the old regime (share of respondents indicating the reason, many choices allowed)	
	The difficulty in writing the calls for tenders so that every bidder would provide adequate quality	0.77
	The need to take quality of service into account	0.73
	The need to avoid bids by "fly-by-night" operators with a high probability of default winning	0.64
	The need to be able to choose small/local/etc firms, at least once in a while	0.90
	Other, please explain	0.27
Q3	Agree with the view that certain types of suppliers were treated favorably in the old regime (share)	0.80
Q4	Agree with the view that in-house cleaning services got a favorable treatment in the old regime (share)	0.71

Q5	Type of supplier that was favored in the old regime	
	(share of respondents indicating the type, many choices allowed)	
	In-house	0.44
	Large national/international firms	0.06
	Regional firms	0.25
	Local firms	0.63
	Favored supplier type varied across municipalities	0.94
Q6	Agree with the view that certain types of suppliers were treated unfavorably in the old regime (share)	0.75
Q7	Type of supplier that was unfavored in the old regime	
	(share of respondents indicating the type, many choices allowed)	
	In-house	0.00
	Large national/international firms	0.73
	Regional firms	0.27
	Local firms	0.07
	Un-favored firm type varied across municipalities	0.93
Q8	Reasons for favoring in-house cleaning units	
	(share of respondents indicating the reason, many choices allowed)	
	Local employment	0.92
	Labor union involvement	0.83
	Political involvement	0.90
	Ease of doing business with them	0.92
	Low quality of private firms' service	0.44
	Financial unreliability of private firms	0.22
	Other, please specify	0.70

Notes: The number of respondents was 22, but not all respondents answered to all questions. The shares reported take the non-response into account. The responses are derived from semi-structured interviews that followed a pre-set protocol and had a common core based on a set of survey questions that were presented to all respondents. The interviewer was however allowed to discuss freely about the auctions with the respondends during the interviews, so the interview protocol was not exactly to the same accross the interviews.

Part C: Details on the discrete choice model (Conditional Logit model)

As we explain in the main text, we use the random utility model of McFadden (1974) to study the choice of the winning bid and to estimate how the price sensitivity of the municipalities changed from the old to new regime.

To describe the model, let the municipalities be indexed by *m*, $m = 1,...,M$, premises to be cleaned by *i*, $i = 1,...,I_m$, and bidders (suppliers) by *j*, $j = 1,...,J_{mi}$. The indirect utility of municipality *m* from choosing bidder *j* to clean building *i* is:

$$U_{mij} = \psi_{mi} + (\eta_1 + \eta_2 \times Scoring_{mi}) \times bid_{mij} + \beta' F_j + q_{mij} + \varepsilon_{mij}, \qquad \text{(A\_eq\_4.1)}$$

where ψ_{mi} refers to the additively separable effects of municipal/procurement/object characteristics (including the format of the auction), bid_{mij} to the bid (price) of supplier *j* for object *i* in municipality *m* (in krona per square meter per day), $Scoring_{mi}$ is an indicator that takes value one if the auction format used in auction *i* of municipality *m* was scoring, F_j to supplier attributes, q_{mij} to 'quality', and ε_{mij} to an error term.

The municipal/procurement/object characteristics, ψ_{mi}, reflect the mean utility that municipality *m* obtains when it has its premises cleaned and the object-specific deviations from the mean. It thus captures all additively separable effects of observable and unobservable municipal characteristics on municipal utility, e.g., regional structure, demographics, income distribution, voter preferences, and propensity to procure services. The term also refers to (un)observable object characteristics, such as the type, size, location, etc. of the object. It captures differences in the indirect utility derived, e.g., from having a clean health center as compared to having clean sports facilities. The assumed additive separability of these effects and the distributional (logit) assumption on the error term (see below) allow us to condition all these effects out in the estimation. The term controls in addition for the additively separable effects on the utility of those characteristics of the procurement event that do not vary over the bidding suppliers, such as whether or not entry to the auction was open, which auction format was used, and whether or not the object was auctioned as a part of a multi-object procurement.

The second term in (A_eq_4.1) specifies the effect of a submitted bid on the choice, with the coefficient measuring the weight given to the bid. The interaction terms allow us to explore whether the weight attached to price is different between the two auction formats (or, in some specifications that we estimate, between the two regimes).

The third term in (A_eq_4.1), F_j, allow us to capture the possibility that there are *supplier-specific*, as opposed to object-specific, quality differences (i.e., ex ante corporate-level

quality differences). For example, a piece of information in the bids through which the suppliers are able to 'differentiate themselves' (besides the price) is the corporate identity of the bidder. This may e.g. convey information about the bidder's experience. To capture this, we use supplier-type dummies or fixed effects in some specifications. We can also alternatively include the number of bids a supplier submits during the entire sample period (to proxy reputation) and the number of bids a supplier submits in a given procurement event (to capture elements of "combinatorial bidding", if any).

The fourth term in (A_eq_4.1), q_{mij}, refers to non-price attributes. It allows for the possibility that municipalities care about the quality of cleaning of *a particular object* for which suppliers are bidding (i.e., ex ante object-level quality differences). The extensive documentation available to us on the technical specifications of the procurements, calls of tenders and the specifics of the bids however suggest that it is likely that there are *no* major ex ante quality differences *at the object-level*. That is, conditional on the corporate identity of the bidders, it is not likely that there are large, ex ante discernible quality differences *between the bids for a specific object*. There is little room for a supplier to differentiate one-self quality-wise, conditional on F_j, suggesting $q_{mij} \approx 0$.

The last term in (A_eq_4.1), ε_{mij}, is a stochastic error term that captures intrinsic randomness in municipality decision making. Given ψ_{mi}, the error term only contains bidder-object specific unobservables. It therefore allows for idiosyncrasies in the decision-making of the procurement bureaucrats. We assume that ε_{mij} was unobservable to bidders and distributed i.i.d. type I extreme value.

Given the above assumptions, and imposing approximation $q_{mij} \approx 0$, the probability that bidder *w* wins in a procurement auction for object *i* organized by municipality *m* is (McFadden 1974):

$$\Pr[y_{mi} = w] = \frac{\exp\{\tilde{U}_{miw}\}}{\sum_{j=1}^{J_{mi}} \exp\{\tilde{U}_{mij}\}} \qquad \text{(A\_eq\_4.2)}$$

where

$$\tilde{U}_{miw} = \psi_{mi} + (\eta_1 + \eta_2 \times Scoring_{mi}) \times bid_{miw} + \beta' F_w$$

$$\tilde{U}_{mij} = \psi_{mi} + (\eta_1 + \eta_2 \times Scoring_{mi}) \times bid_{mij} + \beta' F_j \,.$$

As specified, the model corresponds to the standard Conditional logit model and can be estimated by maximum likelihood (ML).

Part D: Quantile regressions

Table A.4.2 displays the results for four sets (Panels A-C) of reduced form quantile regressions (Koenker and Bassett 1978), estimated using the sample that includes all submitted bids. The dependent variable is, as before, the logarithm of the price of the cleaning service per square meter and day (frequency), adjusted for inflation. To focus on the more relevant left tail of the bid distribution, we report the estimates for the 5th, 10th, 15th, 25th, 50th and 75th percentiles. The reported standard errors are obtained via bootstrapping (100 replications).

Panel A and B show how the bid distributions differ between the old and new regimes, unconditionally and conditionally. In Panel A, we only include the regime indicator. Consistent with the raw data (see Table 1), Panel A shows that the bids are throughout the entire distribution a little higher in the new regime. In Panel B, we use a reduced form specification which includes as controls W = {*Auctions per procurement*, *Size of premises*, *Size of premises*^{2}, *Contract length*, *Extension*, *Frequency*, *Frequency*^{2}}, X = {*School*, *Office*, *Day-care center*, *Unemployment*, *Population density*, *Education*, *Red majority*}, *Trend* (T) and the municipal fixed effects (μ). Panel B shows that conditioning on {W, X, T, μ} matters. The bids of the new regime now appear to be lower in the lower tail and higher in the upper tail of the distribution.

Table A4.2: Characterizing the bid distribution

Panel A: Y=Log(bid), Sample: All submitted bids, Method: Quantile regression

	Percentiles				
	p5%	p10%	p25%	p50%	p75%
Regime (R)	0.060**	0.041*	0.044***	0.047***	0.032*
	(0.030)	(0.022)	(0.015)	(0.018)	(0.017)
Control variables (W, X, T)	No	No	No	No	No
Municipal FE	No	No	No	No	No
Observations	7,427	7,427	7,427	7,427	7,427

Panel B: Y=Log(bid), Sample: All submitted bids, Method: Quantile regression

	Percentiles				
	p5%	p10%	p25%	p50%	p75%
Regime (R)	-0.202	0.022	0.133	0.158***	0.176***
	(0.131)	(0.109)	(0.085)	(0.057)	(0.044)
Control variables (W, X, T)	Yes	Yes	Yes	Yes	Yes
Municipal FE	Yes	Yes	Yes	Yes	Yes
Observations	7,364	7,364	7,364	7,364	7,364

Panel C: Y=Log(bid), Sample: All submitted bids, Method: Quantile regression

	Percentiles				
	p5%	p10%	p25%	p50%	p75%
Regime (R)	-0.113***	-0.076***	-0.054***	-0.036*	-0.047***
	(0.036)	(0.027)	(0.021)	(0.021)	(0.017)
D_inhouse	-0.212***	-0.185***	-0.160***	-0.130***	-0.132***
	(0.031)	(0.021)	(0.017)	(0.015)	(0.017)
D_inhouse x R	0.509***	0.421***	0.289***	0.253***	0.253***
	(0.072)	(0.065)	(0.046)	(0.081)	(0.063)
Control variables (W, X, T)	No	No	No	No	No
Municipal FE	No	No	No	No	No
Observations	7,364	7,364	7,364	7,364	7,364

Notes: The standard errors reported in parentheses are bootstrapped (100 replications). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

To take into account the special role of the inhouse units, we turn to the quantile regressions reported in Panel C. These models include the indicator variable (*D*), which takes value of one if an inhouse unit participated in an auction, as well as its interaction with the regime indicator. The coefficient of *D* measures how the submitted bids differed in the presence of an inhouse unit in the beauty contests of the old regime from the beauty contests in which such a unit did not participate. The key finding is that the indicator *D* obtains a negative and significant coefficient at each percentile that we report. Additional results are 1) that the coefficient of the

indicator *D* reduces in absolute value when we move up in the distribution, suggesting that those firms bidding low reacted more strongly to the presence of an inhouse bidder than those bidding high; and 2) that the interaction between the indicator *D* and the regime indicator R always obtains a positive and significant coefficient larger in absolute value than the direct effect of *D*.

Part E: Characterizing the nature of favoritism

Here we explore why the municipalities were price insensitive in the old regime. To this end, we follow the approach of Bandiera, Prat and Valletti (2009). Applied to our context, their approach suggests that not choosing the lowest bid in the old regime may be evidence of "active waste", which means that the procurement agent directly benefits from the inflated bids. Alternatively, such choice behaviour may mirror "passive waste". Passive waste takes place when for instance poor practices and decisions lead to unintentional, inefficient spending of public resources (Bandiera, Prat, and Valletti, 2009).

We examine, in particular, whether those municipalities that were *less* likely to choose the lowest bid in the old regime were *less* likely to choose a first price auction in the new regime. This revealed preference would be consistent with municipalities actively wasting public resources (Bandiera, Prat, and Valletti, 2009) in the old regime. The opposite would be consistent with passive waste, as it would indicate that when forced to choose between a *First price* and *Scoring* auction, a municipality that in the past often did not award a contract to the lowest bidder (i.e., was price insensitive) now wants to use a competitive first price auction.

We implement this analysis in two steps. In the first step, we form a dummy that takes value one for those auctions of the old regime in which the lowest bid won, and zero otherwise. We use a linear probability model (OLS) and regress this dummy on *W* (object characteristics) and *X* (municipal characteristics, other controls) and the municipal fixed effects, using data

from the old regime only. The estimated municipal fixed effects of this regression provide us with information on which municipalities were particularly price sensitive and likely in the old regime to award the contract to the lowest bidder. The municipal fixed effects are jointly highly significant (F-test = 213.02, p-value < 0.001). In the second step, we use data from the new regime and form an indicator which takes the value one for first price auctions and is zero for scoring auctions. We then regress this indicator on the estimated municipal fixed effects from step one. This regression is also a linear probability model. We display the results from this second step in Table A4.3.

The results show that the coefficient of the municipal fixed effects is positive and significant in Model (1) where we have no other controls and Model (2) where we add contract characteristics. This suggests that those municipalities that were *less* likely to choose the lowest bid in the old regime were also *more* likely to choose a scoring auction in the new regime. This revealed preference is consistent with municipalities "actively wasting" public resources (Bandiera, Prat, and Valletti, 2009) in the old regime.

Model (3) provides some support for the viewpoint that active waste is *unlikely* to mean outright corruption. It shows that when we add the municipal characteristics (*Unemployment*, *Population density*, *Education*, and *Red majority*), the coefficient of the municipal fixed effects remains positive but is no longer significant. This is an interesting finding, because it indicates that our municipal-level control variables are pretty good at capturing the determinants of the choice of the auction format in the new regime. Although not shown in the table, the *Red majority* - indicator, which is one for those municipalities where leftwing parties have a controlling majority in the municipal council, obtains a negative and significant coefficient in Model 3. This is consistent with such municipalities actively not using first price auctions and thus choosing auction formats that allow non-price considerations, such as local employment or union involvement, when picking the winner. The other significant predictor in Model 3 is

the unemployment rate, but it obtains a little surprisingly a positive coefficient. In the estimation sample, the pairwise correlation of *Red majority* with the estimated municipal fixed effects is -0.56 (p-value < 0.001), whereas that of *Unemployment* and the fixed effects is -0.03 (p-value = 0.59).

Table A4.3: Choice between First price and Scoring auctions (New regime)

	OLS Model (1)	OLS Model (2)	OLS Model (3)
Municipal FE	0.438***	0.326**	0.094
	(0.100)	(0.147)	(0.162)
Control variables: Contract characteristics	No	Yes	Yes
Control variables: Municipal characteristics	No	No	Yes
Observations	252	252	252

Notes: The standard errors reported in parentheses are clustered by procurements. *** p < 0.01, ** p < 0.05, * p < 0.10. The dependent variable is an indicator which takes value one for the first price auctions and is zero for the scoring auctions. The municipal FEs are estimated using data from the old regime.

As can be seen from Table A4.1 (see especially Q8), the interview respondents' answers provide further support for the above considerations and particularly for the view that active waste was *unlikely* to mean outright corruption. The four main reasons for the favorable treatment of inhouse units in the old regime were local employment, political involvement, labor union involvement, and ease of doing business with them. These reasons were mentioned by more than four out five respondents.

References used in this Appendix:

Bandiera, O. Prat, A., and Valletti, T. "Active and passive waste in government spending: Evidence from a policy experiment." American Economic Review, Vol. 99 (2009), pp. 1278-1308.

Bergman, M. A. and Lundberg, S. "Tender evaluation and supplier selection methods in public procurement." Journal of Purchasing and Supply Management, Vol. 19 (2013), pp. 73-83.

Koenker, R. and Bassett Jr. G. “Regressions quantiles.” Econometrica, Vol. 46 (1978), pp. 33-50.

McFadden, D. L. “Conditional logit analysis of qualitative choice behavior.” In Frontier in Econometrics, by Paul Zarembka, 105-142. New York: Academic Press, 1974.

Appendix E: Auxiliary material to Section 7

In this Appendix, we provide a summary table of the interview questions used in Section 7.

Table A5.1 Interview evidence (referred to in Section 7)

Q1	Delivered quality of cleaning worse or better in the new regime (1=much worse, ..., 5=much better)	3.26
Q2	Delivered ancillary quality worse or better in the new regime (1=much worse, ..., 5=much better)	3.85
Q3	Renegotiation due to non-delivery/inadequate delivery in	
	Old regime (1=very infrequent, …, 5=very frequent)	1.94
	New regime (1=very infrequent, …, 5=very frequent)	2.23
Q4	Renegotiation due to municipality altering the conditions in	
	Old regime (1=very infrequent, …, 5=very frequent)	2.20
	New regime (1=very infrequent, …, 5=very frequent)	1.82
Q5	Non-performance (any reason) more common in	
	New regime (1=much less common, …, 5=much more common)	3.26
Q6	Early termination of contract due to non-delivery/inadequate delivery in	
	Old regime (1=very infrequent, …, 5=very frequent)	1.36
	New regime (1=very infrequent, …, 5=very frequent)	1.73
Q7	Main problems in delivering cleaning services	
	(share of respondents indicating the reason, many choices allowed)	
	The supplier keeping the schedule (old regime)	0.17
	The supplier keeping the schedule (new regime)	0.18
	The quality of the cleaning (old regime)	0.44
	The quality of the cleaning (new regime)	0.77
	The quality of the cleaning equipment used (old regime)	0.28
	The quality of the cleaning equipment used (new regime)	0.05
	The quality of the personnel (old regime)	0.22
	The quality of the personnel (new regime)	0.50
	Financial issues with the supplier (old regime)	0.17
	Financial issues with the supplier (new regime)	0.23
	General unreliability of the supplier (old regime)	0.22
	General unreliability of the supplier (new regime)	0.27
	Deteriorating performance towards the end of the contract (old regime)	0.28
	Deteriorating performance towards the end of the contract (new regime)	0.32
	Other, please specify (old regime)	0.00
	Other, please specify (new regime)	0.32
Q8	Perceived cost differences between suppliers in	
	Old regime (0 = yes, 1 = no)	0.31
	New regime (0 = yes, 1 = no)	0.33
Q9	If perceived cost differences in the old regime, cost ranking of suppliers	
	(1=lowest cost, …, 5=highest cost)	
	In-house	2.88
	Large national/international firms	2.13
	Regional firms	2.29
	Local firms	2.00
Q10	If perceived cost differences in the new regime, cost ranking of suppliers	

(1=lowest cost, …, 5=highest cost)	
In-house	3.36
Large national/international firms	1.77
Regional firms	1.92
Local firms	2.54

Notes: The number of respondents was 22, but not all respondents answered to all questions. The shares reported take the non-response into account. The responses are derived from semi-structured interviews that followed a pre-set protocol and had a common core based on a set of survey questions that were presented to all respondents. The interviewer was however allowed to discuss freely about the auctions with the respondends during the interviews, so the interview protocol was not exactly to the same accross the interviews.