Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Bidding on price and quality: An experiment on the complexity of scoring rule auctions

Riccardo Camboni*, Luca Corazzini†, Stefano Galavotti‡, Paola Valbonesi§

Abstract. We experimentally study procurement auctions when both quality and price matter. We compare two treatments where sellers compete on one dimension only (price or quality), with three treatments where sellers submit a price-quality bid and the winner is determined by a scoring rule that combines the two offers. We find that, in the scoring rule treatments, efficiency and buyer's utility are lower than predicted. Estimates from a Quantal Response Equilibrium model suggest that increasing the dimension of the strategy space imposes a complexity burden on sellers, so that a simpler mechanism like a quality-only auction may be preferable.

JEL Codes: C90; D44; H57.

Keywords: laboratory experiment; procurement auctions; scoring rule auctions; multi-attribute auctions; complexity.

Acknowledgements: *We are especially grateful to Shachar Kariv and three anonymous Referees for their comments that substantially improved the paper. We thank Patrick*

*Dept. Economics and Management, University of Padova (IT), riccardo.camboni@unipd.it.

†Dept. Economics, University of Venezia "Ca' Foscari" (IT); and Masaryk University Experimental Economics Laboratory, Brno (CZ), luca.corazzini@unive.it.

‡Dept. Economics, Management and Corporate Law, University of Bari (IT), stefano.galavotti@uniba.it.

§Dept. Economics and Management, University of Padova (IT); and National Research University - Higher School of Economics, NRU-HSE (RU), paola.valbonesi@unipd.it.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Bajari, Maria Bigoni, Alessandro Bucciol, Uri Gneezy, Ben Greiner, Marco Guerzoni, Elisabetta Iossa, Gregory Lewis, Antonio Nicolò, Salvatore Nunnari, Fausto Pacicco, Giancarlo Spagnolo, Steve Tadelis, Orestis Troumpounis, Luigi Vena, Andrea Venegoni, Claudio Zoli, participants at the AMEC in St. Petersburg, the European ESA Meeting in Dijion, the ESA World Meeting in Vancouver, the SIOE Conference in Stockholm, and participants at the seminars in Padova and Verona for useful discussions and comments. We are grateful to Joachim Vosgerau for the opportunity to run the experiment at BELSS, Bocconi University, Milan. Iuliana Iuras and Marco Magnani provided excellent research assistance during the experiment. Financial support was provided by Centro Studi di Economia e Tecnica dell'Energia Giorgio Levi Cases, University of Padova, Italy (project n. BERT_EPPR_P14_01); by the PRIN 2017Y5PJ43, funded by the Italian Ministry of Education; and by the Progetto di Eccellenza Fondazione Cariparo 2017, Unipd. The Ethical Committee of the Department of Economics, University of Venice "Ca' Foscari", approved the experimental protocol. All errors are ours.

1 Introduction

In procurement markets, suppliers compete for the right to sell goods or to provide services to a buyer. Usually, the object of the transaction is a good or service that will be delivered in the future and, accordingly, realized from scratch. In many of these cases, a number of valuable attributes of the item to be procured (technical characteristics, delivery lead time, payment conditions, etc.) are contractible, and the buyer's problem is to jointly select a contractor and the contract characteristics, with the goal of obtaining the best *value for money*.

The design of the tender procedure is central to achieving this goal. Two auction mechanisms are usually adopted in practice. In the simplest one, corresponding to a standard first-price auction, the buyer defines the minimal technical requirements in the call for tender and then lets suppliers bid on price only, awarding the contract to the lowest-price seller. Alternatively, the buyer may adopt a scoring rule (or multi-attribute)

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

auction in which participants submit a multidimensional bid comprising a price and a number of non-price attributes; these elements are then mapped, usually in a linear combination fashion, into a *score*, and the supplier that earns the highest score is awarded the contract.

Scoring rule auctions are increasingly used in Europe and in the United States. In Europe, Directive 2014/24/EU on public procurement supports moving away from tenders based on first-price auctions and towards those based on scoring rule auctions (the so called 'most economically advantageous tender', MEAT). In 2021, 57% of all public contracts above the value of 150,000 € were awarded with MEAT (TED, Tender European Daily). In the US, scoring rule auctions have been largely adopted to award highway construction and maintenance: they work as two-dimensional mechanisms, with time to completion as the non-price dimension (Lewis and Bajari, 2011; Gupta et al., 2015). Mechanisms based on price and quality attributes are also adopted in IT services and online freelance auctions.

The increasing popularity of scoring rule auctions is theoretically grounded: Che (1993) and Asker and Cantillon (2008) highlighted the desirable welfare properties of these auctions. The intuition is straightforward: when suppliers are heterogeneous, a scoring rule auction promotes competition, in that it allows suppliers to find the best balance among the various attributes of the offer, and between these attributes and the price offer. Given the relevance of the procurement market, it is then important to test whether this auction mechanism, beyond theoretically appealing, does also perform well in practice. Our conjecture is that a scoring rule auction is arguably a complex strategic environment in that bidders have to reason multidimensionally, and this complexity may prompt behavioral responses that potentially undermine its theoretical superiority. This conjecture is also inspired by the evidence on combinatorial auctions, where the multidimensionality of the bidders' problem has been repeatedly raised as potentially harmful for efficiency (see the discussion at the end of Section 2). Moreover, in examining data on the non-price dimension of scoring rule auctions for highway construction, Lewis and Bajari (2011) observe that "*bidder heterogeneity accounts for more of the variance than contract*

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

heterogeneity" (p. 1201), leaving open the possibility that some relevant behavioral effects may be at work.

While there have been some relevant attempts to study the properties of the scoring rule auctions in the field, the strong heterogeneity across different procurement and institutional settings, as well as the presence of relevant reputational concerns make it often difficult to gather causal and conclusive evidence from real data. To overcome these problems, in this paper we rely on a controlled experiment: by manipulating the dimension of the choice set, we assess how bidders respond to the complexity of the mechanism and, thereby, how alternative auction formats (with varying degrees of complexity) perform in terms of buyer's utility and overall efficiency.

Specifically, we design an experiment with five treatments. Our baseline treatment, named *SRA*, is a scoring rule auction: sellers submit a two-dimensional offer comprising a price and a quality bid, which are then linearly combined according to a publicly announced scoring rule, and the seller whose score is the highest is awarded the contract. Besides, we implement two treatments with one-dimensional bids: one of these, named *FPA*, is a standard first-price auction, where the level of quality is imposed by the buyer, sellers compete on price only, and the seller who submits the lowest price wins the auction. In the other one-dimensional treatment, named *FQA* (which stands for first-quality auction), the buyer announces the price she will pay for the contract, sellers compete on quality only, and the seller who submits the highest bid wins the auction.[1] Finally, the

[1]Though apparently uncommon, *FQA* is not a mere theoretical construct. The European Union Directive on public procurement envisages that *"the cost element may also take the form of a fixed price or cost on the basis of which economic operators will compete on quality criteria only"* (Directive 2014/24/EU, art. 67, second paragraph). Moreover, public calls for research grants often take the form of a competition on quality only: for example, according to the guidelines of the European Research Council, peer reviewers should evaluate research proposals on the basis of excellence as the sole criterion. We are also aware of a few examples in Italy in which this awarding rule has been used in the procurement of care services such as accommodation for asylum seekers and psycho-

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

remaining two treatments lie halfway between *SRA* on one hand, and *FPA* and *FQA* on the other: in these treatments, called *SRA2p* and *SRA2q*, sellers bid on both price and quality, but one of the two bids (price in *SRA2p*, quality in *SRA2q*) is constrained to a binary choice. In all treatments, the exogenous parameters were pin down to maximize the expected utility that the buyer would accrue if sellers were risk neutral and bid according to equilibrium.

Our experimental results highlight the existence of a trade-off between optimality and complexity: while *SRA* is theoretically superior to all other treatments in terms of buyer's utility and total welfare, this is no longer true in the lab. In particular, *FQA* performs as well as *SRA* which, in turn, does not perform better than *SRA2p* and *SRA2q*. Finally, *FPA* has the worst performance, as predicted by the theory.

To shed light on these findings, we then turn to the analysis of bids. Two facts clearly emerge. First, bids tend to be noisier in the scoring rule auctions than in the one-dimensional ones, so that, in the former treatments, there is a larger inefficiency loss due to allocating the contract to the high-cost supplier. Second, the scores associated with the submitted bids are lower than predicted (underbidding) in the scoring rule auctions, whereas the opposite (overbidding) occurs in *FQA* (we observe a slight overbidding, but not significant, also in *FPA*): this is at the origin of the unexpected observed ranking in terms of buyer's utility. We conjecture that these two facts are the results of the suppliers' response to the complexity of the auction mechanisms.

To corroborate this intuition, we fit to our data a structural Quantal Response Equilibrium (QRE) model with two parameters: an error parameter, that measures the degree at which suppliers play suboptimal strategies, and a risk aversion parameter. Across treatments, we obtain very similar estimates for the risk aversion parameter. Estimates for the error parameter are consistent with our intuition: as we move from one-dimensional to two-dimensional treatments, we observe increasing deviations of actual bids from the payoff-maximizing ones. Moreover, across the scoring rule treatments, errors are less pronounced in *SRA2p* and *SRA2q*, where one of the two dimensions is simplified to a binary

pedagogical activities for kids in primary schools.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

choice. After showing how the two elements of the QRE model – risk aversion and error proneness – affect the average bidding behavior and, thereby, the buyer's utility, we finally discuss their implications for market design. To this end, we conduct a counterfactual exercise based on the QRE estimates to show how the market designer should adjust the exogenous parameters in each treatment to optimally accommodate suppliers' behavioral responses. Our analysis suggests that, while it is certainly useful for the buyer to take into account the suppliers' risk attitude when setting the treatment-specific parameters, this element is of secondary importance. Instead, the crucial market design decision is which auction mechanism to implement, as the different degrees of complexity associated with the various mechanisms produce large effects on behavior and outcomes.

The rest of the paper is organized as follows. Section 2 discusses the relevant literature we contribute to, while Section 3 describes our experimental design. Theory and testable predictions are presented in Section 4, and the experimental results are analyzed in Section 5. In Section 6, after relating our experimental findings to a real world example, we first present the QRE approach and then we use these results to perform a counterfactual analysis. Section 7 concludes, drawing some policy implications.

2 Related literature

The theoretical properties of scoring rule auctions were first derived by Che (1993) in a framework in which only one non-price attribute is relevant: he proves that the optimal mechanism is implemented by a quasi-linear scoring rule that under-rewards quality relative to the buyer's true preferences. Asker and Cantillon (2008, 2010) generalize the analysis to a situation where sellers' types are multidimensional, and several non-price attributes matter to the buyer. They also show that, in terms of buyer's utility, the scoring rule auction strictly dominates a price-only auction with minimum quality standards.

A few empirical papers study scoring rule auctions in field (see, e.g., Cameron, 2000, Hyytinen et al., 2018, Kong et al., 2022). The results of these studies provide useful insights on the relative performance of multi-attribute auctions with respect to alternative

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

awarding procedures but, given the peculiarities of the rules and contexts investigated, cannot easily be compared across them and with our experimental results. In this respect, the empirical setting that has more similarities with our experimental design is the one studied by Lewis and Bajari (2011). They compare the scoring rule and first-price auction schemes used by the California Department of Transportation to award more than 1,300 highway construction projects between 2003 and 2008. The quality component in the scoring rule is the number of days to complete the project. They find that projects awarded through the scoring rule auction are more expensive than those awarded with a first-price auction and are completed much sooner. Using a dollar-value estimate of the negative externality to commuters caused by each day of work, they conclude that the users' welfare gain from using a scoring rule instead of a first-price auction largely outweighs the increase in the procurement cost. To fully assess the welfare effects, they then structurally estimate the contractors' cost, assuming optimal behavior at the bidding stage, and conclude that scoring rule auctions generate a significantly larger social welfare than first-price auctions, so they should always be adopted. Moreover, even a policy of small incentives meant to reduce the procurement cost (i.e., a small weight to the quality component in the scoring function) would be welfare improving. Our experimental analysis adds controlled evidence on the actual performance of scoring rule auctions: in line with Lewis and Bajari (2011), our results confirm their superiority over a comparable first-price scheme, though we also show how other simpler mechanisms (notably, *FQA*) may perform as well as a scoring rule auction. Interestingly, when comparing our experimental bids in *SRA* with those collected by Lewis and Bajari (2011), we observe a very similar, large in size, unexplained variability in the quality dimension, which in turn provides supporting evidence of our complexity argument and reassures us about the external validity of our results.

We also contribute to the limited experimental literature on multi-attribute auctions. These papers focus almost exclusively on the performance of the various awarding mechanisms, and, unlike ours, do not deeply analyze the suppliers' side. Chen-Ritzo et al. (2005) run an experiment involving an English reverse auction in which sellers submit

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

three-dimensional bids (price, quality and lead time), and the buyer does not fully disclose how bids are mapped into the score. They find that the three-attribute auction is effective in increasing both the buyer's and the sellers' surplus, although differences are smaller than predicted. Strecker (2010) studies the effect of revealing information in an English auction with three attributes and finds that efficiency (but not buyer's utility) is greater when the scoring rule is fully disclosed than when only limited information is provided to sellers. Bichler (2000) employs an experimental setting that mimics the financial market to assess the performance of three multi-attribute mechanisms – a first-score sealed bid like our *SRA*, a second-score sealed bid, and a first-score open-cry auction – with respect to a single-attribute mechanism. The object of the auction is a call option where the quality element is represented by the volatility of the underlying index or share. He finds that the buyer achieves higher utility in the multi-attribute mechanisms than in the single-attribute one, whereas the level of efficiency is similar.[2]

Finally, our study contributes to the literature exploring how individuals act in complex strategic environments. Auctions are certainly one setting where the issue of complexity is pervasive, especially when multiple items are sold/procured, either sequentially (see Corazzini et al., 2019, and the references therein) or simultaneously (see the survey by Kwasnica and Sherstyuk, 2013). In the context of combinatorial auctions, Kwasnica et al. (2005) refer to the "*computational complexity of the bidders' problem*" as a potential cause of reduction in an auction's efficiency; Kagel et al. (2010) show that suppliers tend to myopically bid on a small number of packages, which may negatively affect efficiency; Scheffel et al. (2012) find that suppliers use simple heuristics to select packages and argue that this approach has to do with cognitive limits in terms of the number of items on which people can simultaneously concentrate. Our paper shows that, even in the apparently simpler context of a single-unit auction, a high degree of complexity, in the form of a multiple number of dimensions on which suppliers are called to think and bid, may

[2]Albano et al. (2018) investigate scoring rule auctions in which quality is exogenously and randomly assigned to each supplier prior to competing. They find that a higher weight attached to quality in the scoring rule increases efficiency.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

affect the auction's performance.

3 Experimental design

3.1 Baseline game and treatments

The baseline game, *SRA*, consists of a procurement scoring rule auction with incomplete information. Two sellers participate in an auction to sell an object to a buyer. The sellers simultaneously place their bids, consisting of two integer numbers: the quality of the object to be sold, denoted by q, and the price at which they are willing to sell it, denoted by p. The submitted quality is constrained to be a number between 0 and 70; the set of admissible prices varies with the submitted quality, and it is constrained by a price cap in the form $p^{\max}(q) = q + 50$.[3] Each seller's bid (q, p) is then mapped into a score s that linearly combines quality and price according to the following *scoring rule*:[4]

$$s(q, p) = 50 + 2q - p. \tag{1}$$

Notice that (1) rewards quality and penalizes price. The coefficients attached to q and p in (1) are set optimally in a sense that will be explained in the next section. The constant term is clearly immaterial, but is added to avoid negative scores. The seller whose score is higher wins the auction, and ties are broken randomly. The winning seller is paid the submitted price p but has to bear the cost of providing the submitted quality q: her monetary payoff is then $m(q, p; \theta) = p - C(q; \theta)$, where the cost function is given by

$$C(q; \theta) = \frac{q^2}{4\theta}, \tag{2}$$

[3]We imposed a price cap to avoid excessive payments. The price cap was set in a way that *(i)* any seller could always make a bid that yields a strictly positive payoff in case of winning, and *(ii)* it was not binding in the equilibrium of our benchmark model.

[4]The choice of a *linear* scoring function (with integer coefficients) was motivated by the desire to keep the *SRA* easily understandable for the experimental subjects. Moreover, real-world multi-attribute auctions usually adopt linear scoring rules.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

and θ, which is idiosyncratic to each seller, identifies the seller's type. On the other hand, the loser of the auction earns nothing. At the beginning of the auction, sellers' types are independently drawn from a discrete uniform distribution with support $\Theta = \{1, 2, 3, \ldots, 10\}$. Each seller observes the realization of her own type but not that of her opponent. Everything else is common knowledge. Notice that the cost function (2) is strictly increasing in q, strictly decreasing in θ (hence, θ can be interpreted as an indicator of the seller's productive efficiency), and convex in both arguments: for given seller's type θ, the marginal cost of quality is increasing; and, for given quality q, the cost difference between two consecutive types gets smaller and smaller as θ increases. The choice of this functional form was motivated by two considerations. First, the convexity with respect to q, an assumption that matches the analysis by Che (1993) and Asker and Cantillon (2008), was necessary to have an interior solution in the derivation of the theoretical equilibrium. Second, (2) has the nice property that it generates a linear equilibrium for our baseline treatment *SRA* (see next section). Given that our main goal is to explore the complexity of *SRA* that stems from the multidimensionality of the bidding decision, it was important to avoid further computational difficulties. Of course, this came at the cost of having non-linear equilibria for the other (simpler) treatments.

Along with the baseline game just described, we consider four additional treatments in which the size and the dimensionality of the sellers' strategy sets are gradually reduced. In two treatments, *FPA* and *FQA*, sellers bid on one dimension only – price in *FPA* and quality in *FQA* – while the other dimension is exogeneously set. Specifically, in *FPA*, sellers are constrained to deliver quality $\bar{q} = 16$ (and to bear the associated cost defined by (2) if they win) and simply submit a price bid. The awarding rule is the same as in *SRA* – the higher-score seller wins the auction –, but since quality is fixed, the winner is simply the seller who submits the lower price. In *FQA*, the buyer commits to pay the price $\bar{p} = 32$ to the winner, and sellers compete on quality only. Since the price is fixed, the seller who offers the higher quality wins the auction (and bears the cost associated with the submitted quality, as defined by (2)). In the remaining two treatments, named *SRA2q* and *SRA2p*, sellers make two-dimensional bids, like in *SRA*, but one dimension -

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

quality in the former treatment and price in the latter - is constrained to a dichotomous choice. Specifically, in *SRA2q* sellers can submit one of two possible quality levels, either $q_L = 9$ or $q_H = 40$, whereas the price bid can be any (integer) value between 0 and $p^{\max}(q)$. In *SRA2p*, the only admissible prices are $p_L = 12$ and $p_H = 65$, whereas any (integer) quality no greater than 70, for $p_L = 12$, and included between 15 and 70, for $p_H = 65$, can be submitted. As in *SRA*, the winner of the auction is the seller whose score, as defined by (1), is higher. The parameters $\bar{q} = 16$ for *FPA*, $\bar{p} = 32$ for *FQA*, $q_L = 9$ and $q_H = 40$ for *SRA2q*, and $p_L = 12$ and $p_H = 65$ for *SRA2p*, have been chosen optimally (see next section).

Throughout the paper, we will often use the term *one-dimensional* auctions to encompass treatments *FPA* and *FQA*; and the term *two-dimensional* auctions (or simply *scoring rule* auctions) to encompass treatments *SRA*, *SRA2q*, *SRA2p*.

3.2 Procedures

Upon arrival, subjects were randomly assigned to a computer terminal, instructions were distributed and read aloud.[5] After reading the instructions, subjects answered a number of control questions to ensure they understood the instructions and the effects of their choices. The experiment started only after all subjects had correctly answered the control questions; when necessary, answers to these questions were explained privately. In each session, subjects participated in 15 consecutive repetitions (or periods) of the game. At the beginning of the experiment, the computer randomly formed four rematching groups of six subjects each. The composition of the rematching groups was kept constant throughout the session. Subjects were randomly and anonymously divided into pairs within their rematching group in every period, and informed that pairs were formed in a way that they would never interact with the same opponent in two consecutive periods.[6]

[5]The English translation of the instructions used in *SRA* are reported in Web Appendix A.

[6]Our rematching protocol implies that, given the size of the sub-group (six subjects), subjects interacted with the same opponent once every five periods, on average. Al-

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Before submitting their final choice(s), subjects could exploit a 'user-friendly' interface to simulate the consequences of their provisional choices in terms of the score associated with that quality/price bid, the cost borne in case of winning, and their earnings. At the end of every period, the outcome of the auction and the subject's earnings were summarized on the screen.

For each treatment, we ran three sessions with 24 subjects each, thus generating 12 independent observations at the rematching group level. The experiment took place at the Bocconi Experimental Laboratory for Social Sciences (BELSS) of Bocconi University, Milan, between December 2017 and January 2018. Most participants were undergraduate students who were recruited by means of the SONA recruitment system (http://www.sona-systems.com/default.aspx) from a pool of around 3000 registered users. The experiment was computerized using the z-Tree software (Fischbacher, 2007). Prices, costs and earnings in the experiment were expressed in tokens converted at an exchange rate of 1 euro per 7 tokens; at the end of the experiment, monetary earnings were paid in cash privately. Subjects started the experiment with a balance of 20 tokens to cover the possibility of losses. On average, subjects earned 14.47 euro for sessions that lasted seventy minutes, including the time for instructions and payments. Before leaving the laboratory, subjects completed a short questionnaire containing questions on their socio-demographics and their perceptions of the experimental task.

4 Theory and predictions

Our experimental results will be compared to the predictions delivered by a benchmark model of risk neutral suppliers and equilibrium behavior.[7] Specifically, we consider a

though this approach is not a perfect stranger protocol, it leaves little room for developing punishment-reward strategies over multiple periods. The rematching protocol was intended to increase the number of independent observations and, therefore, to enhance the statistical power of the non-parametric tests used in the analysis.

[7]The theoretical results are derived in Web Appendix B.1 under the assumption that types and bids are continuous variables, whereas our experimental subjects faced a dis-

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

model in which: *(i)* each seller's utility function coincides with her monetary payoff, and *(ii)* sellers bid according to the (symmetric) Bayes-Nash equilibrium of the auction.

To evaluate the performance of the various treatments in terms of welfare, we set the following utility function for the buyer:

$$u_B(q,p) = \frac{20}{7}q - p. \tag{3}$$

The weights attached to quality and price in (1) are those that maximize the ex-ante expected utility of a buyer with objective function (3), conditional on the sellers playing their equilibrium bidding strategies in a scoring rule auction with linear scoring rule. It is important to stress that the buyer's utility (3) differs from the optimal linear scoring rule (1): in particular, relative to the utility of the buyer, the optimal scoring rule under-rewards quality, a result that is consistent with what already shown by Che (1993). Likewise, the two admissible values for quality in *SRA2q* (price in *SRA2p*) are those that maximize the buyer's ex-ante expected utility, conditional on the sellers bidding their equilibrium strategies in an auction game like *SRA2q* (*SRA2p*) that uses (1) as an awarding rule. Finally, the exogenous value of quality in *FPA* (price in *FQA*) is set to maximize the buyer's ex-ante expected utility, conditional on the sellers bidding according to equilibrium in an auction game like *FPA* (*FQA*).

Figure 1 here

Figure 1a displays the equilibrium scores as a function of θ in the five treatments.[8] Notice that, in all treatments, the equilibrium score is strictly increasing in the seller's

crete setting. The continuous approach allowed us to use calculus and was very useful at the time of optimizing over the exogenous parameters. In Web Appendix B.3, we show that, for *SRA*, the discrete equilibrium coincides with its continuous counterpart; and that, for the other treatments, although there is no perfect coincidence, the equilibrium of the continuous model is a quite accurate approximation of the discrete one.

[8]Although in *FPA* (*FQA*) sellers choose price (quality) only, here and elsewhere we look at their scores as defined by (1). This allows to directly compare bids across treatments.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

type: hence, theoretically, the auction is always won by the seller with the higher θ. Notice also that the equilibrium score is, for all types, highest in *SRA* and lowest in *FPA*. The remaining three treatments – *FQA*, *SRA2q*, and *SRA2p* – lie in between, but the ranking among them is ambiguous: for relatively low types, the equilibrium score of *FQA* is well below that of *SRA2q* and *SRA2p*, but the first is steeper and eventually overtakes the latter two, almost reaching *SRA*. Overall, the equilibrium scores in the five treatments become more concentrated as θ increases.

Figure 1b displays the equilibrium quality bid (remind that quality is fixed in *FPA*). In *FQA*, *SRA2p* and *SRA* – where quality can be set freely – the submitted quality is strictly increasing in the seller's type, but it increases more quickly in *SRA* than in *FQA*. In *SRA2p*, the submitted quality is rather flat for $\theta \leq 6$ and $\theta \geq 7$, but jumps up between $\theta = 6$ and $\theta = 7$. This pattern closely resembles what happens in *SRA2q* (where only $q_L = 9$ and $q_H = 40$ are admissible). Notice, finally that, in *SRA*, the equilibrium score and the equilibrium quality (thereby, also the equilibrium price) are linear in θ, whereas the equilibrium bids are non-linear in the other treatments.

In terms of welfare implications, we will look at the utility of the buyer (as defined by (3)), the payoffs of the sellers, and the total welfare generated, that we measure simply as the sum of the buyer's utility (3) and the winning seller's monetary payoff, yielding $TW(q^w, \theta^w) = \frac{20}{7}q^w - C(q^w; \theta^w)$, where q^w is the quality submitted by the winner of the auction and θ^w is her type.

Table 1 reports, for each treatment, the expected buyer's utility (BU), the expected sellers' payoff (SP), and the expected total welfare (TW, which is simply the sum of BU and SP) associated with the equilibrium bids, expressed in relative terms with respect to the maximum total welfare achievable (see the rows with heading 'Pred.'). It turns out that the buyer is better off with the scoring rule treatments: specifically, in equilibrium, *SRA* generates the highest BU, followed by *SRA2p*, *SRA2q*, *FQA*, and *FPA*.[9] When

[9]The ranking in terms of buyer's utility described here refers to our specific model. However, the superiority of scoring rule auctions with respect to one-dimensional auctions is a fairly general result. It is obvious that an optimally designed *SRA2p* (*SRA2q*) is

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

looking at sellers, the ranking largely reverses: the *SP* is higher in the one-dimensional treatments than in the two-dimensional ones. Finally, the ranking along TW fully mirrors that in terms of BU.

Concerning total welfare, it is also useful to look at its determinants. As the expression for TW suggests, there are two dimensions that jointly affect efficiency. The first is *cost efficiency*: whatever level of quality is delivered, the object should be produced at the lowest possible cost. The second dimension is *quality efficiency*: the level of quality delivered by the winning seller should be such that its marginal cost is equal to its marginal benefit (to the buyer); since the marginal benefit of quality is constant and equal to 20/7, and the marginal cost is $q/(2\theta)$, the efficient level of quality when the object is delivered by a type-θ seller is $q^{\text{EFF}}(\theta) = (40/7)\theta$. It is immediate to see that, in equilibrium, all treatments are cost-efficient: in fact, since scores are strictly increasing, the object is always assigned to the low-cost seller (i.e., the seller with the higher θ in the pair). Therefore, the inefficiency that characterizes all treatments (TW is always less than one) is due to quality inefficiency: in particular, as shown by Figure 1b, with some exceptions for $\theta < 3$, the submitted quality falls short of its efficient level (denoted by FB in the Figure).

The welfare rankings outlined above can be understood in light of the differences in the strategy spaces across treatments. Intuitively, in *SRA* and, to a lesser extent, in *SRA2q* and *SRA2p*, sellers have more flexible strategies at their disposal, as they can leverage on both quality and price to compete in the auction. In particular, a seller whose cost for quality is high can still be competitive by pairing a low quality bid with a low price. This choice is not possible in treatments where sellers bid on price or quality only. As a

certainly better for the buyer than an optimally designed *FQA* (*FPA*). Moreover, Che (1993) shows that the optimal mechanism for the buyer is indeed a *quasi-linear* scoring rule auction: hence, our *SRA*, which is the best among the scoring rule auctions with *linear* scoring rule, is not necessarily the optimal mechanism, but cannot be too far from optimality. On the other hand, the ranking between *SRA2p* and *SRA2q*, and the ranking between *FQA* and *FPA* are both sensitive to the primitives of the model.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

result, competitive pressure is stronger in the two-dimensional treatments: this increases efficiency and favors the buyer to the detriment of sellers. Notice also that the shape of the cost function (2) is at the origin of the poor performance of *FPA*: in fact, with quality fixed exogenously, the convexity of the cost function generates cost differences that get larger as θ decreases. As a consequence, the competitive pressure from low to high types is extremely weak, negatively affecting both the the buyer's utility and the total welfare.

We summarize the main theoretical predictions in the following statements.

H.1 Buyer's Utility. In equilibrium, the ranking with respect to expected buyer's utility is as follows: $SRA \succ SRA2p \succ SRA2q \succ FQA \succ FPA$.

H.2 Sellers' Payoff. In equilibrium, the ranking with respect to expected sellers' payoff is as follows: $FQA \succ FPA \succ SRA \succ SRA2q \succ SRA2p$.

H.3 Total Welfare. In equilibrium, the ranking with respect to expected total welfare is as follows: $SRA \succ SRA2p \succ SRA2q \succ FQA \succ FPA$.

All treatments are cost efficient, whereas no treatment is quality efficient. Hence, the ranking in terms of quality efficiency mirrors the ranking in terms of total welfare. Quality inefficiency is due to the fact that, in all treatments and with some exceptions for low types, the submitted quality is below the efficient level.

H.4 Bids. In all treatments, the equilibrium score functions are strictly increasing in the type parameter θ. For all types, the score is maximal in *SRA* and minimal in *FPA*; *FQA*, *SRA2q* and *SRA2p* lie in between.

5 Experimental results

The experimental results are presented in two steps. First, we concentrate on the welfare generated by our five treatments, looking separately at the buyer's utility, the sellers' payoff and the total welfare. Next, we analyze bidding behavior by looking at the observed score as defined by (1). The non-parametric tests presented in the next pages are based on twelve independent observations (at the rematching group level) per treatment. Moreover,

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

when looking at differences across treatments over all periods, we will also discuss results from the bootstrap-based methodology developed by List et al. (2019) to test multiple null hypotheses simultaneously in experimental settings with multiple treatments. Results of the bootstrap-based methodology will be identified by the acronym MHT. In the parametric analysis, we properly account for dependency of observations over repetitions by either clustering standard errors, or introducing random effects at the rematching group level. All regressions pool data from the five treatments and use FQA as a baseline.

5.1 Welfare

The top part of Table 1 shows the descriptive statistics for the observed BU, SP, and TW (see the rows with heading 'Avg.'), the corresponding predicted levels ('Pred.') and the results from a (two-sided) Wilcoxon signed-rank test for the null hypothesis of equality between observed and predicted levels. In the Table, the observed loss in total welfare (WL, which is equal to $1-TW$), is decomposed into the two components of cost inefficiency (CI) and quality inefficiency (QI). Specifically, QI is computed by taking, for each pair and each period, the difference between the level of total welfare that would have been generated if the winner had submitted its (type-specific) efficient quality level and the actual level of total welfare observed in the auction. On the other hand, CI, which captures the welfare loss due to inefficiently assigning the contract to the high-cost seller, is obtained as the difference between the first-best welfare level and the level of total welfare that would have been generated if the winner of the auction had submitted its efficient quality level.

All these measures are expressed in relative terms: namely, for each pair and in each period, we divide each measure by the level of welfare associated with the first-best allocation (i.e., the level of overall surplus that would have been generated if the good had been awarded to the low-cost seller and this seller had provided the efficient quality level). All measures are then averaged by period and by rematching group.

Table 1 here

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

When compared to theory, data seem to display a rather clean qualitative difference between one-dimensional and two-dimensional treatments. In terms of BU, indeed, *FPA* (+5.3%) and, especially, *FQA* (+11.1%) significantly outperform their theoretical prediction, whereas the scoring rule auctions underperform (the difference between the predicted and the observed buyer's utility is negative and significant in *SRA* and *SRA2p*, it is negative but not significant in *SRA2q*). As a result, *FQA*, which ranked fourth theoretically (see prediction H.1), is the best mechanism for the buyer in the lab (66.5% of the potential surplus), followed by *SRA* (62.5%), *SRA2p* (62.0%), and *SRA2q* (59.7%).

A similar pattern, but in the opposite direction, is detectable when looking at the suppliers' side: *SP* is lower than predicted in the one-dimensional treatments (−6.6% in *FPA*, −7.8% in *FQA*), while it is aligned with theory in the scoring rule auctions.

The overperformance of *FQA* and the underperformance of the scoring rule auctions in terms of BU passes on to TW, producing a similar ranking: *FQA*, which ranked fourth theoretically (see prediction H.3), is the most efficient mechanism in the lab (83.4% of the potential surplus), followed by *SRA* (82.3%), *SRA2p* (80.5%), and *SRA2q* (80.1%).

Notice that, overall, the observed differences across *FQA* and the scoring rule auctions are small for both BU and TW, whereas *FPA* is by far the worst treatment along both welfare measures.

Looking at the determinants of the observed welfare loss, we see that no treatment is fully cost efficient: in this respect, the best treatment is, by far, *FQA*, where the efficiency loss due to awarding the contract to the high-cost seller amounts to 1.6% of the potential welfare. This percentage is higher for *FPA* (5.4%) and for the scoring rule treatments (5.7% in *SRA2q*, 6.3% in *SRA*, 7.1% in *SRA2p*). With respect to quality inefficiency, instead, the observed ranking fully obeys the theoretical one, with *FPA* being, not surprisingly, the worst treatment (recall that, in this treatment, quality was fixed). Finally, it is interesting to notice that the negative difference between observed and predicted total welfare recorded in the scoring rule auctions is essentially due to cost inefficiency, as the observed quality inefficiency is aligned with the predicted one.[10]

[10]Notice also that the fraction of the observed welfare loss attributable to cost ineffi-

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Table 2 here

To assess the statistical validity of these preliminary observations, Table 2 reports parametric results of the determinants of our welfare measures. Column (1) confirms that, as far as BU is concerned, the theoretical ranking among treatments (prediction H.1) is partially upset in the lab: while *FPA* yields the lowest utility to the buyer, as predicted (the pairwise differences between *FPA* and all other treatments are negative and highly significant - in all cases, $p < 0.001$ according to both parametric tests and MHT), *FQA* generates at least as much buyer's utility as the scoring rule auctions: we do not document significant differences between *FQA* and *SRA*, a positive and significant difference between *FQA* and *SRA2q* ($p = 0.008$), and a positive and marginally significant difference between *FQA* and *SRA2p* ($p = 0.060$). Across the scoring rule auctions, differences are not significant.

Columns (3) of Table 2 focusses on *SP*. Notice that, while theory suggests sellers' payoff should be higher in the one-dimensional treatments than in the two-dimensional ones (prediction H.2), our experimental results detect few differences across treatments: the only marginally significant (positive) difference is the one between *SRA2q* and *FQA* ($p = 0.077$), whereas all the other pairwise comparisons are not statistically significant.

The results regarding TW essentially replicate those concerning BU. Column (5) shows that *FPA* is the least efficient treatment, as all the pairwise differences between *FPA* and the other treatments are negative and highly significant (in all cases, $p < 0.001$ according to both parametric tests and MHT). As for *FQA*, we detect no significant difference with respect to *SRA*, positive and marginal significance with respect to *SRA2p* ($p = 0.081$), positive and significant difference with respect to *SRA2q* ($p = 0.039$). No significant differences are observed across the three scoring rule auctions.

Column (7) of Table 2 reports parametric results concerning CI. Results confirm that *FQA* is the most cost-efficient treatment: in fact, all the pairwise comparisons between *FQA* and the other treatments are negative and highly significant (according to MHT:

ciency is much lower in the one-dimensional treatments (9.6% in *FQA* and 13.8% in *FPA*) than in the two-dimensional ones (28.6% in *SRA2q*, 35.6% in *SRA*, 36.4% in *SRA2p*).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

$p < 0.001$ with respect to *FPA*, $p = 0.016$ with respect to *SRA2q*, $p = 0.032$ with respect to *SRA2p* and $p = 0.029$ with respect to *SRA*). We find no significant difference across the scoring rule mechanisms and between any scoring rule auction and *FPA*. The differences in *QI* across treatments are reported in column (9). In this respect, *FPA* is the most inefficient treatment (all the pairwise differences between *FPA* and any other treatment are statistically significant; according to MHT: $p < 0.001$ with respect to *FQA*, *SRA2q* and *SRA2p*; $p = 0.003$ with respect to *SRA*). Instead, there is no significant difference among the other treatments.

All the previous parametric results remain qualitatively unchanged if a linear time trend is added (see columns (2), (4), (6), (8) and (10)).[11]

Below, we summarize the main results concerning welfare.

R.1 Buyer's Utility. *FQA* performs as good as *SRA*, and better than *SRA2q* and *SRA2p*; no differences are detected among *SRA2q*, *SRA2p*, and *SRA*; *FPA* is the worst treatment. The buyer's utility is above its predicted level in *FQA* and *FPA*, and below it in the scoring rule auctions.

R.2 Sellers' Payoff. No remarkable differences are detected across treatments. The sellers' payoff is aligned with its predicted level in the scoring rule auctions, and below it in *FQA* and *FPA*.

R.3 Total Welfare. The ranking in total welfare mirrors the ranking in buyer's utility. With respect to cost inefficiency, *FQA* is the least inefficient treatment,

[11]The main results concerning *BU*, *SP* and *TW* are confirmed when, to account for the effects of subjects' experience, we focus on the last five periods of the experiment only. In particular: with respect to *BU*, *FPA* is still the worst treatment (in all the pairwise differences with the other treatments, $p < 0.001$); with respect to *SP*, all pairwise differences are confirmed. The only marginally significant (positive) differences are between *SRA* and *FQA* ($p = 0.078$), and between *SRA* and *SRA2q* ($p = 0.070$); with respect to *TW*, we do not detect any difference between *FQA* and, respectively, *SRA*, *SRA2p* and *SRA2q*, whereas *TW* is significantly lower in *FPA*. See Web Appendix C.2 for details.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

while no significant differences are documented among the remaining treatments. With respect to quality inefficiency, *FPA* is the worst treatment, while no significant differences are documented among the remaining treatments. The total welfare is above its predicted level in *FQA* and below its predicted level in the scoring rule auctions and, to a lesser extent, in *FPA*.

5.2 Bidding behavior

In what follows, we analyze the submitted bids to gain a better understanding of the determinants of the puzzling evidence on buyer's utility and total welfare. For ease of comparison, we focus on the scores associated with the sellers' bids (the corresponding variable is named *score*). We first assess differences in the bidding behavior across treatments. We then examine how the scores depart from equilibrium using two statistics: the percentage difference between observed and predicted score (denoted as *score_diff*); and the quartile coefficient of dispersion of the bids made by sellers with the same θ (the corresponding variable is denoted *score_qcd*).[12] It is worth noticing that *score_qcd* overcomes the comparability issues associated with other standard dispersion measures. Indeed, *score_qcd* accounts for differences across treatments in the equilibrium relationship between the score and the seller's type, θ (Figure 1a).

The bottom part of Table 1 reports descriptive statistics on *score*.[13] Three main facts stand out. First, unlike what is theoretically predicted (see prediction H.4, Section 4), the average submitted score is highest in *FQA*, followed by *SRA2p* and *SRA2q*. *SRA* is ranked fourth, closely followed by *FPA*. Second, we observe overbidding in the one-dimensional treatments (in particular, the average score in *FQA* is 4.4% higher than predicted) but

[12]Specifically, *score_diff* = (*score* − predicted score)/predicted score; *score_qcd* = $\sum_{\theta=1}^{10}(1/10)[Q_3(s|\theta) - Q_1(s|\theta)]/[Q_3(s|\theta) + Q_1(s|\theta)]$, where $Q_1(s|\theta)$ and $Q_3(s|\theta)$ are the first and the third quartile of the observed frequency of score s submitted by type-θ sellers. For *score* and *score_diff*, observational units refer to the per period measures built at the bidder level.

[13]A pictorial representation can be found in Web Appendix C.1.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

underbidding in the two-dimensional ones (in particular, the average score in *SRA* is 14.4% below its predicted level). A (two-sided) Wilcoxon signed-rank test confirms that *score_diff* is significantly different from zero for all treatments except *FPA*. Third, for given type θ, scores in *SRA* are, on average, 2.7 times more volatile than in *FQA*; in general, the scoring rule auctions exhibit higher dispersion than the one-dimensional treatments.

Table 3 here

Table 3 reports parametric results on differences across treatments and determinants of *score* and of *score_diff*. The baseline specification in column (1) confirms that the submitted score is highest in *FQA* (for all the coefficients of the treatment dummies: $p < 0.001$). We find a nonsignificant difference between *SRA2p* and *SRA2q*, while both these treatments are associated with a higher score than *SRA* (between *SRA2p* and *SRA*, $p = 0.027$; between *SRA2q* and *SRA*, $p = 0.024$). Finally, we do not detect any significant difference between *SRA* and *FPA*.[14] Column (2), which includes the seller's type θ among the regressors, shows that, in all treatments, higher types tend to submit higher scores (in all cases, $p < 0.001$).[15] Finally, column (3) controls for treatment-specific linear time

[14]According to MHT, all the differences between *FQA* and the other treatments remain significant (with respect to *FPA*, $p < 0.001$; with respect to *SRA2q*, $p = 0.002$; with respect to *SRA2p*, $p = 0.010$; and with respect to *SRA*, $p = 0.003$). Moreover, $p = 0.004$ for the difference between *FPA* and *SRA2q*, and $p = 0.003$ for the difference between *FPA* and *SRA2p*.

[15]From column (2), we can also see that that the observed score function is steeper in *FQA* than in the scoring rule auctions, something that is in line with the theory (see Figure 1a). To see this, notice that, in column (2), *SRA*, *SRA2p*, and *SRA2q* have positive coefficients, whereas their interactions with θ are negative. Therefore, we can use the estimates to determine for which types the difference between *FQA* and any of the scoring rule auctions becomes significant. We find that the submitted score in *FQA* is: *(i)* above the score in *SRA* for $\theta \in [3, 10]$; *(ii)* above the score in *SRA2p* for $\theta \in [4, 10]$; and *(iii)* above the score in *SRA2q* for $\theta \in [5, 10]$.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

trends. We detect a positive and significant time pattern in the scoring rule auctions (in all cases, $p < 0.001$) and in *FPA* ($p = 0.002$).[16] Even after controlling for the type parameter and the linear trend, the score in *FQA* remains higher than in *SRA* and *FPA* (in both cases, $p < 0.001$). Moreover, we find a nonsignificant difference between *SRA2p* and *SRA2q*, while scores in these two treatments are higher than in *SRA* ($p = 0.018$ between *SRA2p* and *SRA*; $p = 0.013$ between *SRA2q* and *SRA*). Finally, we find a significantly lower score in *FPA* than in *SRA* ($p = 0.001$).

Table 3 also reports the results on *score_diff*. In the baseline model in column (4), we find a positive deviation (overbidding) of 4.41% in *FQA* ($p < 0.001$) and significant underbidding in *SRA2q* (-5.95%), *SRA2p* (-9.48%), and *SRA* (-14.43%) (in all cases, $p < 0.001$). No significant difference between observed and predicted scores is documented in *FPA*. Column (5) confirms that the previous results remain significant even after controlling for sellers' types: in *SRA* and *SRA2p*, observed scores are significantly below their predicted levels for all type parameters (while, in *SRA2q*, this occurs for $\theta \in [1, 7]$). Instead, in *FQA*, the overbidding is significant for $\theta \in [4, 10]$, and the degree of overbidding increases with θ. For example, a supplier of type $\theta = 5$ is associated with a positive deviation of 3.96% in *FQA* ($p = 0.001$) and negative deviations of 14.85% in *SRA* ($p < 0.001$), 10.18% in *SRA2p* ($p < 0.001$), and 6.53% in *SRA2q* ($p < 0.001$). Controlling for the linear trend in column (6) does not affect the results in *FPA* or *FQA*, but reduces the magnitude of the underbidding that characterizes the scoring rule auctions.[17]

[16] All estimates are reported in Web Appendix C.2.

[17] The significant time pattern detected in the scoring rule treatments may reveal some learning effect. As a robustness check, we replicate all the regressions focusing on the last five periods of the experiment, when the (treatment-specific) trend coefficients are no longer significant. For the dependent variable *score*, we find that the differences between *FQA* and the other treatments persist and are strongly significant, and they depend on the type parameter θ: as θ increases, the score in *FQA* approaches and then exceeds the *SRA*'s score. In particular, we detect that the score in *FQA* is: *(i)* below the score in *SRA* for $\theta \in [1, 3]$, *(ii)* not significantly different for $\theta \in [4, 5]$, and *(iii)* above the score in *SRA* for $\theta \in [6, 10]$. Similar results are obtained when *FQA* is compared with *SRA2p*

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

To assess differences across treatments in bid dispersion, we ran MHT on the variable *score_qcd*. Specifically, for each treatment, subgroup and type, we derived the quartile coefficients of dispersion. Then, the MHT has been run by using 12 independent observations per treatment obtained by averaging the quartile coefficients of dispersion at the subgroup level. The results show that bids in *SRA* are significantly more volatile than in the one-dimensional treatments (*FQA* vs. *SRA*: $p = 0.002$; *FPA* vs. *SRA*: $p = 0.034$), and that bids in *FQA* are significantly less volatile than in the scoring rule treatments (*FQA* vs. *SRA2q*: $p = 0.002$; *FQA* vs. *SRA2p*: $p < 0.001$). We also find that bids in *FPA* are significantly more volatile than in *FQA* ($p = 0.013$). Finally, bids in *SRA* exhibit a significantly higher dispersion than in *SRA2q* ($p = 0.025$). All the remaining pairwise comparisons between treatments yield non significant results.

As a final step in the analysis of bids, we run two robustness checks.[18] First, we check whether the feedback information on the subject who won the auction affects bidding decisions in the following period. To this end, we replicate the regressions in columns (2) and (5) of Table 3 adding a dummy that is equal to 1 if the subject won the auction in the previous period and 0 otherwise, as well as corresponding interactions with the treatment dummies. Results suggest that winning the auction in the previous period does not exert any significant effect on the submitted score, nor on its percentage difference from the predicted level. Second, our parametric results on the submitted scores may hide a significant heterogeneity in individual behaviors. We address this point by estimating, for each subject in *SRA*, the (individual-specific) parameters of the linear bidding function: $s(\theta) = a + b\theta$. We focus on *SRA* because it is the only treatment in which the theoretical bidding function is indeed linear in the type θ (specifically: $s(\theta) = 52 + 2\theta$). We find that the mean of the estimates of a is much lower than the predicted value, while the mean

and *SRA2q*. For the dependent variable *score_diff*, we still find a significant underbidding in *SRA* (-7.54%; $p < 0.001$) and in *SRA2p* (-5.09%, test results: $p < 0.001$) and a significant overbidding in *FQA* ($+4.22\%$; $p < 0.001$). Adding the type parameter θ does not qualitatively alter these results. See Web Appendix C.2 for details.

[18]The results of these robustness checks can be found in Web Appendix C.3 and C.4.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

of the estimates of b is slightly higher than predicted. Taken together, the estimated individual bidding functions almost always produce underbidding in the score.

Below, we summarize the main results concerning bidding behavior.

R.4 Bids. In all treatments, the score increases with the type parameter θ. The submitted scores are highest in *FQA*, while reach the lowest level in *SRA* and *FPA*. There is overbidding in *FQA* and underbidding in the scoring rule auctions. Scores are noisier (i.e., they present higher dispersion) in the scoring rule auctions than in the one-dimensional treatments.[19]

6 Discussion and structural analysis

The crucial result of our experiment is that, compared to what is predicted by our benchmark model of equilibrium with risk neutral sellers, the two-dimensional treatments perform significantly worse, both in terms of buyer's utility and total welfare, whereas essentially the opposite occurs for the one-dimensional treatments: both *FQA* and *FPA* overperform in terms of buyer's utility and *FQA* overperforms also in terms of total welfare. As a consequence, the observed rankings (see R.1 and R.3) are partially different from what expected (see H.1 and H.3), with *FQA* doing (at least) as good as the scoring rule treatments. The analysis of bidding behavior, summarized in result R.4, sheds light on the reasons behind these results. In fact, the comparison of observed bids across treatments shows two interesting facts:

Fact I. Noisy Bidding: bids are more noisy in the two-dimensional treatments than in the one-dimensional ones. This explains why, even though in all treatments the low-cost seller submits *on average* a higher score, cost inefficiency, which arises whenever the high-cost seller wins the auction, is larger in the two-dimensional auctions.

[19] Interestingly, the underbidding that we detect in the scoring rule auctions is the result of a higher-than-predicted quality accompanied by an even stronger upward adjustment of the submitted price. For more on the analysis of the price-quality combination in the scoring rule treatments, see Web Appendix C.5.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

This explains why the two-dimensional treatments have lower-than-predicted total welfare;

Fact II. Overbidding/Underbidding: in the two-dimensional treatments, the submitted scores are lower than what theoretically predicted, whereas the opposite occurs in the one-dimensional auctions.[20] This explains why the two-dimensional (one-dimensional) treatments perform worse (better) than predicted in terms of buyer's utility.

Facts I and II above seem to reflect a significant difference in bidders' behavior between one-dimensional and two-dimensional treatments: this leads us to suspect that the bidding behavior could be somewhat related to the degree of complexity of the auction. Intuition suggests that choosing price and quality simultaneously is a more complex task than choosing one dimension only. Besides, when a two-dimensional bid is to be made, the choice is arguably easier when, on one dimension, only two markedly different alternatives are available, as is the case in treatments *SRA2q* and *SRA2p*. According to this intuition, the five treatments considered in our experiment are characterized by different levels of complexity: treatments with one-dimensional choice (*FQA* and *FPA*) are the least complex, *SRA* is the most complex, and *SRA2q* and *SRA2p* – treatments with two-dimensional choice, one of which is binary – lie in between. This intuition is corroborated by the observation of the subjects' response times in the experiment. Table 1 shows, for every treatment, the average time elapsed before a subject submitted her bid in a generic period of the experiment. The difference in the response time between one-dimensional and two-dimensional treatments is remarkable. Moreover, among the scoring rule auctions, *SRA* required more time to answer than *SRA2p* or *SRA2q*.[21]

[20] Remind that, for *FPA*, the slight overbidding detected is not significant.

[21] According to MHT, all pairwise comparisons are highly significant: for the difference between *FPA* and *FQA*, $p = 0.010$; for the difference between *SRA2q* and *SRA2p*, $p = 0.002$; for the remaining pairwise comparisons, $p < 0.001$. A parametric analysis that includes treatment dummies as covariates leads to the same conclusions. Results are available upon request.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Interestingly, the large variability of the bids that we record in our scoring rule treatments (Fact I), especially in *SRA*, is aligned with what documented by Lewis and Bajari (2011) when analyzing the (real-world) scoring rule auctions used to award highway maintenance contracts in California. There are strong similarities between Lewis and Bajari's and our experimental setting, which favor the comparability of the results: only one quality attribute (measured, in their paper, by the time to complete the work), a linear scoring rule, a considerable weight attached to the quality component both in the scoring rule and in the buyer's objective function, convex cost functions for quality provision that do not cross for different bidders' types. Furthermore, our experimental findings regarding the comparison between *FPA* and *SRA* are consistent with theirs: although the buyer pays a (slightly) lower price in the first-price auction, the increase in quality obtained using a scoring rule auction is such that the net effect on the buyer's utility is positive and substantial. More importantly for the scope of the present study, the particularly noisy bidding behavior that we detect in *SRA* – which is at the heart of the underperformance of this format with respect to theory and that we attribute to the complexity of the bidding task – is in line with their findings: regressing the quality bids in *SRA* on the cost parameter and on bidders' fixed effects, we find that around 28% of the overall variance remains unexplained. This number is remarkably close to what we obtain by replicating, this time on the dataset used by Lewis and Bajari, a similar regression of the quality bid, including contract and bidder fixed effects to account for other time-invariant unobserved characteristics: with their data, 30% of the overall variance of quality choices remains unexplained.[22] These similarities reassure us about the external validity of our results.[23]

[22]The results of this comparative analysis are in Web Appendix D.1. We are indebted to Gregory Lewis and Patrick Bajari for sharing their data and codes.

[23]In addition, several studies show that bidding strategies in laboratory settings are not only consonant with the main predictions of theoretical (equilibrium) models, but also well resemble what observed in real-world contexts (see chapter 9 of Lusk and Shogren, 2007, Betz et al., 2017, and the references therein).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

6.1 A Quantal Response Equilibrium model

In this section, we conjecture and test the hypothesis that the level of complexity of the auction scheme affects bidders' behavior in that it increases their propensity to submit suboptimal bids. To this end, we consider a structural model that explicitly envisages and measures this propensity: the Quantal Response Equilibrium (QRE) introduced by McKelvey and Palfrey (1995). The QRE has been successfully used to model non-equilibrium behavior in experimental auctions. In a QRE model, the assumption that a player always chooses her best response to the opponent's strategy is replaced by a probabilistic choice function tuned by an error parameter: the probability of playing a suboptimal strategy is strictly positive, but it depends on the (relative) payoff associated with it. In other words, an individual is more likely to make an error that determines a small loss (relative to the payoff-maximizing strategy) than an error that causes a big loss. Applied to our context, a (symmetric) QRE is an array of probabilities $\pi = \{\pi_{\theta,b}\}_{\theta\in\Theta, b\in B_\theta}$, where each element $\pi_{\theta,b}$ – the probability that a type-θ seller bids b – is the solution to the following (logistic) equation:

$$\pi_{b;\theta} = \frac{\exp[U_S(b;\theta|\pi)/\mu]}{\sum_{b\in B_\theta}\exp[U_S(b;\theta|\pi)/\mu]}, \tag{4}$$

where $U_S(b;\theta|\pi)$ is the expected utility of a type-θ seller when she bids b conditional on the fact that the other seller bids according to π, B_θ is the set of (admissible) individually rational bids for that type, and $\mu \geq 0$ is the error parameter – the higher μ, the higher the probability the seller makes a bid that yields a relatively low payoff.[24]

We also allow for possible departures from risk neutrality by considering a Constant Relative Risk Aversion (CRRA) utility function for sellers. In particular, the utility of a seller who wins the auction, is paid a price p, and delivers a quality q is equal to $u_S(p,q;\theta) = (1-r)^{-1}\left[p - C(q;\theta)\right]^{1-r}$, where $r \geq 0$ is the Arrow-Pratt coefficient of relative risk aversion, and $C(q;\theta)$ is given by (2).

[24]Hence, the QRE is the solution of a system of $\sum_{\theta\in\Theta}|B_\theta|$ equations. Notice that, when $\mu \to 0$, the QRE model boils down to the standard Bayes-Nash equilibrium.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Our conjecture is that the value of the error parameter μ – which measures the degree of departure from optimal bidding – increases with the complexity of the auction mechanism at hand;[25] and that this, when coupled with risk aversion, is able to rationalize the two observed facts concerning bidding behavior that were outlined before. In fact, notice that, with respect to our benchmark model of equilibrium with risk neutral sellers, in all treatments: *(i)* risk aversion ($r > 0$) induces sellers to submit a higher score (overbidding);[26]; *(ii)* the presence of payoff-sensitive errors ($\mu > 0$) not only adds noise to the bidding behavior (by definition), but also should cause a reduction in the average score submitted by sellers (underbidding); moreover, the higher μ, the larger the degree

[25]We do not investigate the underlying cognitive process that leads individuals to make more suboptimal choices as the problem at hand becomes more complex. Perhaps this may be the result of a trade-off between cognitive effort and the quality of the decision: the individual decides the amount of cognitive effort to devote to a task by weighing the extra-cost of additional effort with its extra-benefit in terms of (expected) improvement of the solution to the task. As a result, when a task is highly demanding in terms of cognitive costs, the decision maker can (optimally) decide to stop thinking about it when a satisfactory, but not necessarily the best, solution has been identified.

[26]That risk aversion leads bidders to bid more aggressively in standard independent private value single-unit first-price auctions is a well known result (see, e.g. Krishna, 2009). This result immediately applies to our one-dimensional auctions: with respect to the case of risk neutrality, sellers submit a lower price in *FPA*, a higher quality in *FQA*, i.e., a higher score in both cases. It is easy to see that risk aversion leads sellers to overbid on the score also in *SRA*: this is achieved by submitting a lower price than under risk neutrality, whereas the equilibrium quality bid is unaffected by the sellers' risk attitude (see Liu et al., 2012). In Web Appendix B.4, we show that overbidding on the score carries over also to treatments *SRA2q* and *SRA2p*. Notice that risk aversion has been identified in the literature as the leading explanation, though not without controversy, for the overbidding phenomenon that is predominantly observed in independent private value single-unit first-price auctions (for a survey, see Kagel, 1995; Kagel and Levin, 2011).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

of underbidding. We do expect this underbidding effect because, looking at the shape of the expected payoff of a seller in the equilibrium (under risk neutrality), it appears that deviations above the payoff-maximizing score are more costly than deviations below.[27]

Hence, if an increase in the complexity of the auction mechanism increases the tendency of sellers to make errors in a QRE fashion (i.e., it increases μ), then this would result in noisier bids in the more complex, two-dimensional treatments (Fact I); and, provided that sellers are risk averse, in a transition from overbidding to underbidding as the complexity of the mechanism increases (Fact II).

Table 4 here

Table 4, panel (A), collects the estimates of the two free parameters of the QRE model: the error parameter μ and the coefficient of relative risk aversion r. For computational reasons, estimations were performed after grouping bids into bins. In particular, the space of admissible bids were divided into 3-unit disjoint intervals (parallelograms with base and height equal to 3 for *SRA*), and, to each observation belonging to a certain interval (parallelogram), the central value was assigned. We then computed the QRE strategies for each type-θ and each pair (r, μ), and, finally, performed standard maximum-likelihood techniques to select the set of parameters that best fits the experimental data.[28] Observe that the estimates for the risk-aversion parameter r are always greater than 0: hence, sellers seem indeed to be averse to risk. Importantly, these estimates are very similar across treatments, which is reassuring about the appropriateness of our randomization protocol and the sensibleness of the QRE model. On the other hand, consistently with our intuition on the complexity of treatments, there are significant differences in the error parameter: the simplest, one-dimensional treatments have smaller values of μ with respect to the two-dimensional ones. Hence, submitted bids are closer to the best responses and

[27] In Web Appendix B.2, we show this graphically for *FQA* and *SRA*.

[28] The binning methodology and the maximum-likelihood procedure are described in Web Appendix D.2 and D.3. For *FPA* and *FQA*, we were able to estimate the model without bins, obtaining very similar results: $r = 0.65$, $\mu = 0.72$ for *FPA*; $r = 0.67$, $\mu = 0.38$ for *FQA*.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

less noisy in the former than in the latter (Fact I). Notice, moreover, that, within the scoring rule auctions, the estimated value of μ is lower in the simpler *SRA2q* and *SRA2p*.[29]

Figure 2 here

Figure 2 displays, for treatments *FQA* and *SRA*, the median (type-specific) score predicted by the QRE model and compares it with the observed bids. For ease of reference, the figure also reports the Bayes-Nash equilibrium under risk neutrality (*BNE RN*) and under risk aversion (*BNE RA*). These equilibria have been obtained numerically exploiting our QRE model: by switching off both parameters, we obtained *BNE RN*; by switching off the error parameter and setting the risk aversion parameter equal to the value inferred from our experiment, we obtained *BNE RA*.

Notice that the QRE model correctly predicts what observed in the experiment and summarized in Facts I and II: noisier bids in *SRA* than in *FQA*; underbidding in the former, overbidding in the latter. Figure 2 also confirms that, relative to the equilibrium under risk neutrality, risk aversion leads sellers to bid more aggressively (*BNE RA* is always above *BNE RN*). On the other hand, error proneness operates in the opposite direction, reducing the submitted scores: in fact, the QRE predicted median score (which includes both risk aversion and errors) is equal or below *BNE RA*. When errors are relatively frequent, as it occurs in *SRA*, the second effect prevails and sellers eventually underbid on the score.

[29]While there are several studies applying a QRE model to experimental auction data (see, among others, Hortaçsu and Bajari, 2005, and Camerer et al., 2016), as far as we know there are no contributions focusing on scoring rule auctions. Therefore, we cannot directly compare the estimated parameters in *SRA*, *SRA2q*, and *SRA2p* with existing results. Nevertheless, Goeree et al. (2002) represents a comparable benchmark for our *FPA* treatment: their estimate of the risk aversion parameter is similar to ours ($r = 0.56$ vs. $r = 0.68$), whereas their error parameter is smaller ($\mu = 0.26$ vs $\mu = 0.78$). Most likely, this discrepancy is due to the difference in the size of the strategy sets between their setting (only 7 admissible bids) and ours (up to 16 possible binned bids).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Observe that the estimates of the parameter μ come from auction games with different strategy spaces. To facilitate comparability across treatments, we also constructed an ex-post measure of departure from rationality, denoted by η, that is less sensitive to the details of the underlying game, being directly built on the relative payoffs predicted by the QRE model. Specifically, $\eta \in [0, 10]$ is computed as (the sum over types of) the average quadratic deviation between the utility associated with the QRE strategy and the maximum utility achievable (the one obtainable by playing the best response strategy with probability one), normalized by the latter. In symbols:

$$\eta = \sum_{\theta=1}^{10} \sum_{b \in B_\theta} \widehat{\pi}_{b;\theta}(b) \cdot \left(\frac{U_S(b; \theta | \widehat{\pi}) - U_S(b^*(\theta); \theta | \widehat{\pi})}{U_S(b^*(\theta); \theta | \widehat{\pi})} \right)^2,$$

where $\widehat{\pi}_{b;\theta}$ is the probability that a type-θ seller bids b, as predicted by the QRE model, and $b^*(\theta)$ is her utility-maximizing bid. Hence, in the expression for η, sellers are indeed risk averse (with the risk aversion parameter arising from our maximum likelihood estimation) and they play their own (noisy) QRE strategy (with the error parameter arising from our maximum likelihood estimation). Loosely speaking, η is a sort of 'money-left-on-the-table' measure, as it captures how much of the potential utility the subject gives up, on average, by using a suboptimal strategy. Now, the ranking across treatments in terms of η supports our starting intuition even more cleanly than when we look at μ: the value of η in *FPA* and *FQA* is much lower than it is in the two-dimensional treatments; moreover, it is higher in *SRA* than it is in *SRA2q* and *SRA2p*.[30]

Finally, we checked how well the QRE model fits our experimental data. Looking at the values of the log-likelihood function may be problematic, as different treatments involve different games with different strategy spaces. To overcome this problem, we follow

[30] One may wonder whether these differences in rationality disappear once subjects learn 'how to play'. To address potential learning dynamics, we re-estimated the baseline QRE model, restricting our attention to the last 5 periods, where no trend was parametrically observed. Results are fully consistent with the estimates on the full sample. See Web Appendix D.4 for details.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Camerer et al. (2016) and adopt a normalized measure of relative fit that is invariant to the dimension of the strategy set. This measure, denoted ϕ_M, which is analogous to a *Pseudo-*R^2, compares the value of the log-likelihood in the estimated model with two extreme models: the first is an ideal 'clairvoyant' model in which each (type-dependent) bid is played with a probability exactly equal to the observed relative frequency; the second is a purely random model in which, for every type-θ, each (individually rational) strategy is played with equal probability. In the first four treatments, *FPA*, *FQA*, *SRA2p* and *SRA2q*, ϕ_M is comparatively high: with respect to a purely random choice, our model explains between 73% (in *FPA*) and 83% (in *FQA*) of the observed bids. The value of ϕ_M reduces to 38% for the *SRA* treatment. This latter value is broadly in line with the results obtained by Camerer et al. (2016) in a (richer than ours) QRE model applied to maximum value experimental auctions.[31]

6.2 A counterfactual analysis

The structural analysis just presented has identified two sources of deviations from the benchmark equilibrium framework with fully rational and risk-neutral bidders: risk aversion, captured by the parameter r, and error-proneness, captured by the parameter μ. Remarkably, while the estimated value of the former parameter is quite similar across treatments, the latter is increasing with the complexity of the auction mechanism.

A natural question then arises: what should a buyer do if, in the auctions' design stage, she took into account the behavioral responses elicited by the different awarding mechanisms? To address this question, we use the predictions delivered by our QRE

[31]We also estimated a more flexible QRE model with one additional parameter meant to capture a possible bias in the sellers' perception of the marginal cost of quality. Interestingly, with this additional parameter the relationship between complexity of the mechanism and error proneness is preserved (the ranking in μ is unaltered), but the fit of the model improves for treatment *SRA* (the other treatments are unaffected). Hence, it seems that, in *SRA*, sellers underestimate the convexity of the cost function. This *augmented* QRE model is presented in Web Appendix D.6.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

model to perform a counterfactual analysis that is meant to derive the optimal values of the parameters controlled by the market designer. These design parameters are: for *SRA*, the weight a attached to quality in the (linear) scoring rule $s(q,p) = aq - p$; for *FPA* (*FQA*), the fixed quality $\bar{q}$ (the fixed price $\bar{p}$); for *SRA2q* (*SRA2p*), the two admissible quality levels q_L and q_H (the two admissible price levels p_L and p_H). Remind that, to pin down the values of these parameters to be used in the experiment, we maximized the expected buyer's utility, assuming that sellers are risk neutral and bid according to the Bayes-Nash equilibrium of the auction game. Here we do the same, but this time assuming that sellers behave as predicted by the QRE model, with the parameters observed in the lab (actually, for the risk aversion parameter, given that we observed minor differences across treatments, we used a common intermediate value, $\hat{r} = 0.67$, for all treatments).

In Table 4, panel (B), the row with heading *QRE opt* reports the optimal design parameters obtained from this counterfactual exercise together with the associated expected *BU* and *TW*. For ease of reference, the Table also reports the values of *BU* and *TW* predicted by the QRE model with the design parameters used in our experiment (*QRE*), as well as those associated with the equilibrium under risk neutrality (*BNE RN*) and risk aversion (*BNE RA*). Table 4, panel (B), provides insights on the effects of risk aversion and error proneness. Starting from *BNE RN*, sellers' risk aversion is good news for the buyer: as we have seen before, risk aversion leads sellers to bid more aggressively (overbidding), namely to submit a lower price in *FPA*, a higher quality in *FQA*, a price-quality combination that yields a higher score in the scoring rule auctions. Overbidding greatly increases the buyer's utility (and also total welfare) in all treatments, without affecting the rankings. On the other hand, the tendency of sellers to make (payoff-sensitive) errors has the opposite effect: sellers, on average, bid more conservatively, and this reduces the buyer's utility (and total welfare): this is clearly seen from comparing the configuration *BNE RA* and *QRE*. Notice how the negative effect on buyer's utility is much stronger in the scoring rule auctions (and especially in *SRA*), where deviations from the payoff-maximizing bids are larger.

In light of these considerations, what should then the buyer do? The last row shows

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

how the treatment-specific design parameters should be adjusted in order to maximize the buyer's expected utility. In particular: (i) in *FPA*, the fixed quality should be increased (from 16 to 20); (ii) in *FQA*, the fixed price should be increased (from 32 to 38); (iii) in *SRA*, the weight attached to quality in the (linear) scoring rule should be increased (from 2 to 2.4); (iv) in *SRA2q*, the low quality should be increased (from 9 to 12); (v) in *SRA2p*, the low price should be decreased (from 12 to 6) and the high price should be increased (from 65 to 84).

The intuition behind these results is clear: the buyer takes advantage of the sellers' risk aversion by accommodating their tendency to bid more aggressively, which, in turn, yields to a higher submitted quality. This improves her utility (and also increases total welfare), while the ranking across treatments is essentially preserved.

Notice, however, that the extra-utility the buyer can gain by optimally setting the design parameters in each treatment is of small magnitude (between +1% in *SRA2q* and +9% in *SRA2p*). Hence, these results suggest that the very crucial decision for the buyer is which auction mechanism to implement. Through this choice, the market designer can limit the negative effect of the sellers' tendency to make errors, which, as we showed, is largely influenced by the complexity of the mechanism adopted. In particular, our results show that, even after optimizing over the design parameters to incorporate bidders' behavioral biases, a simpler mechanism like *FQA* is no worse for the buyer (if not better) than the more complex scoring rule auctions.

7 Concluding remarks

In this paper we experimentally studied the problem of a buyer who wants to procure an item for which both price and quality matter. We considered five auction mechanisms, that differ in their intrinsic trade-off between theoretical performance and bidding complexity. In the simplest mechanisms, *FPA* and *FQA*, sellers bid on one dimension only (price or quality, respectively), while the other dimension was set by the experimenter. In the most complex mechanism, *SRA*, sellers chose both price and quality, bids were

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

(linearly) combined into scores, and the seller with the higher score won the auction. We also considered two scoring rule auctions of intermediate complexity, where the sellers' choice set on one dimension (price in *SRA2p*, quality in *SRA2q*) was only binary.

Our experimental results show that the theoretical ranking across treatments is partially upset in the lab. In particular, in contrast to the theoretical predictions, *FQA*'s performance is (at least) as good as that of *SRA2p*, *SRA2q* and *SRA*, both in terms of buyer's utility and social welfare.

The analysis of bidding behavior shed light on this puzzling evidence. In particular, we observe a quite clean difference in behavior between one-dimensional and two-dimensional auctions: first, bids tend to be more noisy in the latter than in the former, and, second, the scores associated with the submitted bids are lower than predicted (underbidding) in the scoring rule auctions, whereas the opposite (overbidding) occurs in *FQA* and, marginally, in *FPA*. We conjecture that these two facts are the results of the suppliers' response to the complexity of the auction mechanisms. In fact, by estimating a structural QRE model of bidding behavior (that allows for risk averse sellers), we found strong evidence in favor of a positive relationship between complexity of the mechanism and bidders' proneness to make suboptimal bids. In the more complex two-dimensional auctions, the greater tendency of bidding away from the best response generates more noisy behavior which undermines the efficiency of the allocation and produces, on average, more conservative bidding which reduces the utility of the buyer. Finally, we ran a counterfactual analysis that suggests that the observed unexpected ranking across treatments is not due to a suboptimal choice of the design parameters: in fact, even after fine-tuning the treatment-specific parameters to account for the bidders' behavioral responses elicited by the different awarding mechanisms, buyer's utility in the scoring rule auctions is still no greater than in *FQA*.

Hence, our paper suggests that, in general, a market designer should seriously take into account the potential distortions triggered by the complexity of the market mechanism adopted. More sophisticated and theoretically superior mechanisms may perform worse than expected because of the complexity burden they impose on market participants. In

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

particular, the choice among mechanisms with different degrees of complexity is likely to be even more crucial than the design of the fine details of each mechanism.

In the specific context of procurement auctions, our results confirm the main findings of Lewis and Bajari (2011): in a price-quality setting, when the quality attribute of the contract is particularly important, both on the demand side and on the supply side, letting bidders compete both on quality and price through a scoring rule auction is certainly better for the procurer than running a first-price auction, even though this is certainly a more straightforward mechanism. However, our paper also shows that, in this case, it might be optimal to let bidders compete only on the non-price attribute, avoiding the complexity associated with a two-dimensional mechanism. Hence, the choice of the attributes on which sellers are called to compete is a key one and should be evaluated case by case. In fact, any additional attribute in the scoring rule, which, *per se*, has a positive pro-competitive effect, also involves a complexity cost, which may offset the former. Further research is required to assess how the solution to this trade-off changes as the relative importance of the various (price and non-price) attributes varies.

References

Albano, G. L., A. Cipollone, R. Di Paolo, G. Ponti, and M. Sparro (2018). Scoring rules in experimental procurement. *CESIEG WP 10*, 1–26.

Asker, J. and E. Cantillon (2008). Properties of scoring auctions. *RAND Journal of Economics 39*(1), 69–85.

Asker, J. and E. Cantillon (2010). Procurement when price and quality matter. *Rand Journal of Economics 41*(1), 1–34.

Betz, R., B. Greiner, S. Schweitzer, and S. Seifert (2017). Auction format and auction sequence in multi-item multi-unit auctions: An experimental study. *Economic Journal 127*(605), F351—-F371.

Bichler, M. (2000). An experimental analysis of multi-attribute auctions. *Decision Support Systems 29*(3), 249–268.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Camerer, C., S. Nunnari, and T. R. Palfrey (2016). Quantal response and nonequilibrium beliefs explain overbidding in maximum-value auctions. *Games and Economic Behavior 98*, 243–263.

Cameron, L. J. (2000). Limiting buyer discretion: Effects on performance and price in long-term contracts. *American Economic Review 90*(1), 265–281.

Che, Y. K. (1993). Design competition through multidimensional auctions. *RAND Journal of Economics 24*(4), 668–680.

Chen-Ritzo, C. H., T. P. Harrison, A. M. Kwasnica, and D. J. Thomas (2005). Better, faster, cheaper: An experimental analysis of a multiattribute reverse auction mechanism with restricted information feedback. *Management Science 51*(12), 1753–1762.

Corazzini, L., S. Galavotti, and P. Valbonesi (2019). An experimental study on sequential auctions with privately known capacities. *Games and Economic Behavior 117*, 289–315.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics 10*(2), 171–178.

Goeree, J. K., C. A. Holt, and T. R. Palfrey (2002). Quantal response equilibrium and overbidding in private-value auctions. *Journal of Economic Theory 104*(1), 247–272.

Gupta, D., E. M. Snir, and Y. Chen (2015). Contractors' and agency decisions and policy implications in a+b bidding. *Production and Operations Management 24*(1), 159–177.

Hortaçsu, A. and P. Bajari (2005). Are structural estimates of auction models reasonable? evidence from experimental data. *Journal of Political Economy 113*(4), 703–741.

Hyytinen, A., S. Lundeberg, and O. Toivanen (2018). Design of public procurement auctions: Evidence from cleaning contracts. *Rand Journal of Economics 49*(2), 398–426.

Kagel, J. H. (1995). Auctions: A survey of experimental research. *Handbook of Experimental Economics, Princeton University Press*, 501–586.

Kagel, J. H. and D. Levin (2011). Auctions: A survey of experimental research, 1995–2010. *Handbook of Experimental Economics, Vol. 2, Princeton University Press*, 563–637.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Kagel, J. H., Y. Lien, and P. Milgrom (2010). Ascending prices and package bidding: A theoretical and experimental analysis. *American Economic Journal: Microeconomics 2*(3), 160–85.

Kong, Y., I. Perrigne, and Q. Vuong (2022). Multidimensional auctions of contracts: An empirical analysis. *American Economic Review 112*(5), 1703–36.

Krishna, V. (2009). *Auction theory*. Academic press.

Kwasnica, A., J. O. Ledyard, D. Porter, and C. DeMartini (2005). A new and improved design for multiobject iterative auctions. *Management Science 51*(3), 419–434.

Kwasnica, A. M. and K. Sherstyuk (2013). Multiunit auctions. *Journal of Economic Surveys 27*(3), 461–490.

Lewis, G. and P. Bajari (2011). Procurement contracting with time incentives: Theory and evidence. *Quarterly Journal of Economics 126*(3), 1173–1211.

List, J. A., A. M. Shaikh, and Y. Xu (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics 24*, 773–793.

Liu, S., J. Li, and D. Liu (2012). Multi-attribute procurement auctions with risk averse suppliers. *Economics Letters 115*(3), 408–411.

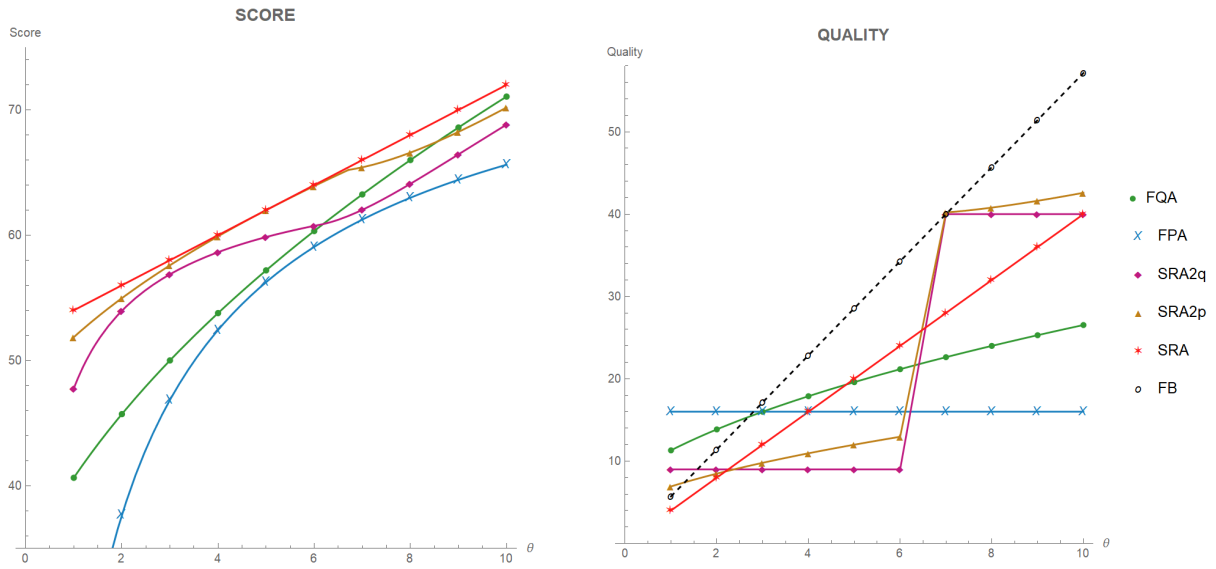
Lusk, J. L. and J. F. Shogren (2007). *Experimental Auctions: Methods and Applications in Economic and Marketing Research - Quantitative Methods for Applied Economics and Business Research*. Cambridge University Press.

McKelvey, R. D. and T. R. Palfrey (1995). Quantal response equilibria for normal form games. *Games and economic behavior 10*(1), 6–38.

Scheffel, T., G. Ziegler, and M. Bichler (2012). On the impact of package selection in combinatorial auctions: An experimental study in the context of spectrum auction design. *Experimental Economics 15*(4), 667–692.

Strecker, S. (2010). Information revelation in multiattribute english auctions: A laboratory study. *Decision Support Systems 49*(3), 272–280.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288



(a) Score (b) Quality

Figure 1: Equilibrium score and quality as a function of θ. FB denotes the first-best quality level.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Table 1: Descriptive statistics.

		FPA	*FQA*	*SRA2q*	*SRA2p*	*SRA*
BU	Avg.	0.43 (0.20)	0.66 (0.09)	0.60 (0.18)	0.62 (0.18)	0.62 (0.22)
	Pred.	0.38 (0.16)	0.55 (0.03)	0.63 (0.08)	0.70 (0.07)	0.71 (0.02)
	p	0.006	0.002	0.158	0.012	0.008
SP	Avg.	0.18 (0.09)	0.17 (0.05)	0.20 (0.15)	0.18 (0.15)	0.20 (0.61)
	Pred.	0.25 (0.09)	0.25 (0.04)	0.20 (0.04)	0.18 (0.03)	0.20 (0.02)
	p	0.003	0.002	1.000	0.530	0.136
TW	Avg.	0.61 (0.19)	0.83 (0.07)	0.80 (0.13)	0.80 (0.14)	0.82 (0.45)
	Pred.	0.62 (0.19)	0.80 (0.06)	0.83 (0.12)	0.87 (0.09)	0.91 (0.00)
	p	0.002	0.002	0.004	0.004	0.002
CI	Avg.	0.05 (0.08)	0.02 (0.03)	0.06 (0.09)	0.07 (0.10)	0.06 (0.09)
QI	Avg.	0.34 (0.20)	0.15 (0.06)	0.14 (0.10)	0.12 (0.10)	0.11 (0.45)
WL	Avg.	0.39 (0.20)	0.17 (0.07)	0.20 (0.13)	0.19 (0.14)	0.18 (0.45)
Obs.		180	180	180	180	180
score	Avg.	52.0 (16.5)	60.5 (12.1)	56.5 (11.9)	56.4 (13.2)	53.9 (12.0)
	Pred.	51.5 (14.6)	57.7 (9.6)	59.8 (5.9)	62.1 (5.7)	62.7 (5.9)
score_diff	Avg.	0.01 (0.16)	0.04 (0.08)	−0.06 (0.16)	−0.09 (0.18)	−0.14 (0.16)
	p	0.530	0.002	0.004	0.002	0.002
score_qcd		0.067	0.040	0.070	0.082	0.107
resp. time	Avg.	26.8 (20.8)	34.7 (23.9)	60.1 (30.4)	70.6 (32.1)	96.8 (39.9)
Obs.		1080	1080	1080	1080	1080

Notes. 'Avg.' is the observed mean of that variable, 'Pred.' is the predicted value (computed from the actual type realizations observed in the lab), *p* is the *p*-value of a (two-sided) Wilcoxon signed-rank test for the null hypothesis of equality between observed and predicted value, *resp. time* is the response time in seconds. Standard errors in parenthesis. The non-parametric tests are confirmed by GLS regressions (with robust standard errors) in which the dependent variable is the difference between the observed and the predicted level of that measure and the controls include a constant term and the treatment dummies.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Table 2: Welfare: parametric analysis.

	BU		*SP*		*TW*		*CI*		*QI*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
FPA	-0.235^{***}	-0.216^{***}	0.010	0.027	-0.224^{***}	-0.189^{***}	0.038^{***}	0.044^{***}	0.187^{**}	0.144^{***}
	(0.023)	(0.040)	(0.015)	(0.019)	(0.016)	(0.031)	(0.007)	(0.012)	(0.016)	(0.028)
SRA2q	-0.067^{***}	-0.144^{***}	0.035^{*}	0.093^{**}	-0.033^{**}	-0.051^{*}	0.041^{***}	0.065^{***}	-0.008	-0.014
	(0.025)	(0.044)	(0.020)	(0.042)	(0.016)	(0.029)	(0.008)	(0.016)	(0.013)	(0.021)
SRA2p	-0.045^{*}	-0.096^{**}	0.016	0.031	-0.029^{*}	-0.065^{***}	0.055^{***}	0.078^{***}	-0.026^{*}	-0.013
	(0.024)	(0.048)	(0.018)	(0.046)	(0.017)	(0.024)	(0.010)	(0.016)	(0.015)	(0.024)
SRA	-0.040	-0.098	0.029	0.035	-0.011	-0.063	0.047^{***}	0.062^{***}	-0.036	0.002
	(0.026)	(0.061)	(0.046)	(0.122)	(0.033)	(0.076)	(0.008)	(0.015)	(0.033)	(0.081)
Trend	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Constant	0.665^{***}	0.647^{***}	0.169^{***}	0.165^{***}	0.834^{***}	0.813^{***}	0.016^{***}	0.022^{***}	0.150^{***}	0.166^{***}
	(0.010)	(0.016)	(0.006)	(0.009)	(0.006)	(0.012)	(0.002)	(0.005)	(0.005)	(0.012)
Obs.	900	900	900	900	900	900	900	900	900	900
Wald $-\chi^2$	100.87	147.76	4.06	16.54	188.67	284.86	101.12	126.45	150.25	244.47
$p > -\chi^2$	0.000	0.000	0.398	0.056	0.000	0.000	0.000	0.000	0.000	0.000

Notes. This table reports estimates (robust standard errors in parentheses) from GLS random effect models accounting for dependency within rematching group. *FPA*, *SRA2q*, *SRA2p* and *SRA* are treatment dummies (*FQA* is the baseline tretment). *Trend* includes a linear time trend and trend-treatment interactions. Significance levels are denoted as follows: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

Table 3: Score: parametric analysis.

	score			*score_diff*		
	(1)	(2)	(3)	(4)	(5)	(6)
FPA	-8.523^{***}	-8.650^{***}	-9.636^{***}	-0.036^{**}	0.046^{**}	0.034
	(1.157)	(1.249)	(1.361)	(0.017)	(0.020)	(0.021)
SRA2q	-4.021^{***}	2.988^{**}	-1.976	-0.104^{***}	-0.123^{***}	-0.206^{***}
	(1.157)	(1.252)	(1.363)	(0.017)	(0.020)	(0.022)
SRA2p	-4.072^{***}	2.741^{**}	-2.137	-0.139^{***}	-0.166^{***}	-0.247^{***}
	(1.157)	(1.255)	(1.360)	(0.017)	(0.020)	(0.021)
SRA	-6.626^{***}	1.482	-5.328^{***}	-0.188^{***}	-0.205^{***}	-0.318^{***}
	(1.157)	(1.246)	(1.355)	(0.017)	(0.020)	(0.022)
θ		3.870^{***}	3.871^{***}		0.009^{***}	0.009^{***}
		(0.093)	(0.088)		(0.001)	(0.001)
$FPA \times \theta$		0.166	0.174		-0.015^{***}	-0.015^{***}
		(0.132)	(0.124)		(0.002)	(0.002)
$SRA2q \times \theta$		-1.251^{***}	-1.215^{***}		0.004^{*}	0.004^{**}
		(0.131)	(0.124)		(0.002)	(0.002)
$SRA2p \times \theta$		-1.235^{***}	-1.238^{***}		0.005^{**}	0.005^{**}
		(0.131)	(0.124)		(0.002)	(0.002)
$SRA \times \theta$		-1.397^{***}	-1.372^{***}		0.003^{*}	0.004^{**}
		(0.130)	(0.123)		(0.002)	(0.002)
Trend	NO	NO	YES	NO	NO	YES
Constant	60.500^{***}	39.188^{***}	38.903^{***}	0.044^{***}	-0.005	-0.010
	(0.818)	(0.889)	(0.966)	(0.012)	(0.014)	(0.015)
Obs.	5400	5400	5400	5400	5400	5400
Wald $-\chi^2$	61.61	6031.89	7405.95	158.64	470.14	1237.72
$p > -\chi^2$	0.000	0.000	0.000	0.000	0.000	0.000

Notes. This table reports estimates from two-way linear random effects models accounting for both potential individual dependency over repetitions and dependency within rematching group. θ is the seller's type. The other remarks of Table 2 apply.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288

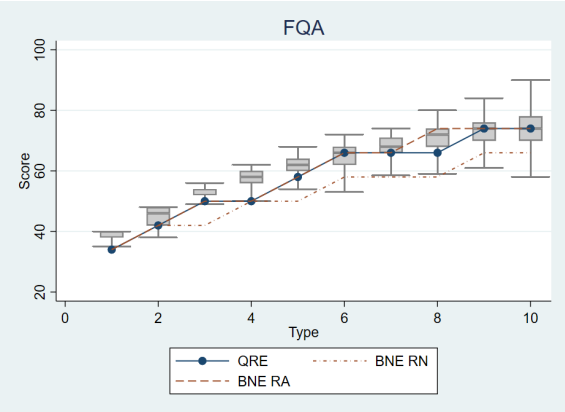
Table 4: QRE: estimates (panel A) and counterfactual analysis (panel B).

(A)	*FPA*	*FQA*	*SRA2q*	*SRA2p*	*SRA*
r	0.68	0.68	0.70	0.66	0.65
μ	0.78	0.42	0.86	1.04	1.22
η	0.389	0.422	0.958	1.277	1.564
	(0.029)	(0.138)	(0.105)	(0.197)	(0.066)
LL	−355.82	−261.97	−467.42	−311.57	−1362.39
ϕ_M	0.73	0.83	0.74	0.76	0.38

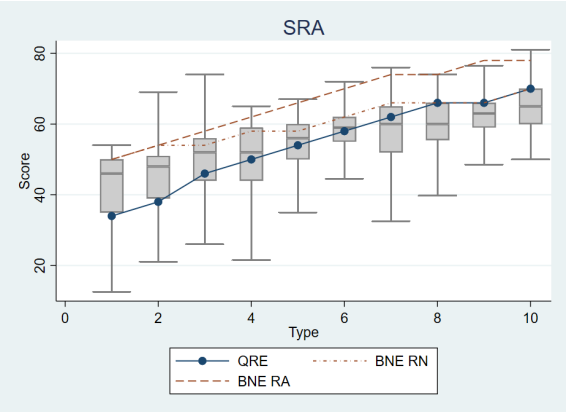
(B)	*FPA*			*FQA*			*SRA2q*			*SRA2p*			*SRA*		
	$\bar{q}$	*BU*	*TW*	$\bar{p}$	*BU*	*TW*	(q_L, q_H)	*BU*	*TW*	(p_L, p_H)	*BU*	*TW*	a	*BU*	*TW*
BNE RN	16	18.4	34.7	32	28.9	44.0	$(9, 40)$	34.5	49.9	$(12, 65)$	38.0	51.8	2	39.6	53.9
BNE RA	16	30.5	34.7	32	41.3	49.2	$(9, 40)$	41.4	50.3	$(12, 65)$	45.9	53.2	2	47.3	54.2
QRE	16	25.4	34.0	32	38.4	47.8	$(9, 40)$	36.3	48.8	$(12, 65)$	32.6	45.6	2	34.3	47.1
QRE opt	20	25.9	39.0	38	39.3	50.0	$(12, 40)$	36.7	51.7	$(6, 84)$	35.6	49.9	2.4	35.4	50.7

Notes. Panel (A) reports: maximum likelihood estimates of the parameters of the QRE model (r and μ); the relative utility loss predicted by the estimated model (η, standard errors in parenthesis); the value of the log-likelihood function (LL); a comparable measure of goodness of fit (ϕ_M). Panel (B): *QRE* corresponds to the Quantal Response Equilibrium with the design parameters used in the experiment, with coefficient of risk aversion set at $\hat{r} = 0.67$ in all treatments, and with the estimated treatment-specific error parameters (see panel (A)). *BNE RA* is obtained by switching off the error parameter, while keeping the risk aversion parameter set at $\hat{r} = 0.67$. *BNE RN* is obtained by switching off both the error and the risk aversion parameters. *QRE opt* is analogous to *QRE* but with the design parameters that maximize *BU*.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01288



(a) *FQA*



(b) *SRA*

Notes. For each type: the gray box comprises scores in the 2nd and 3rd quartile of the observed distribution, the dark-gray segment in each box is the median observation, the two vertical gray lines extend up to 1.5 times the interquartile range. *BNE RN* and *BNE RA* are the Bayes-Nash equilibrium under risk neutrality and under risk aversion, respectively.

Figure 2: Scores in *FQA* and in *SRA*: observed vs. predicted.