

Mitigating Skewed Bidding for Conference Paper Matching

Inbal Rozenzweig¹, Reshef Meir¹, and Nicholas Mattei²

¹Technion — Israel Institute of Technology

²Tulane University

April 30, 2022

Abstract

The explosion of conference paper submissions, especially in artificial intelligence and related disciplines, has underscored the need to improve many aspects of the peer review process, especially the matching of papers and reviewers. Recent work argues that the key to improve this matching, is to modify aspects of the *bidding phase* itself, to ensure that the set of bids over papers is balanced, and in particular to avoid ‘orphan papers’ with no bids.

We have developed a flexible bidding platform to test adaptations to the bidding process. Using this platform, we performed a field experiment during the bidding phase of a medium-size international workshop that compared two bidding methods. We further examined via controlled experiments on Amazon Mechanical Turk various factors that affect bidding, and in particular the order in which papers are presented [Cabanac and Preuss, 2013; Fiez et al., 2020]; and information on paper demand [Meir et al., 2021].

Our results suggest that several simple adaptations, that can be added to any existing platform, have the potential to significantly reduce the skew in bids, thereby improving the allocation for both reviewers and conference organizers.

1 Introduction

Academic peer review of papers and grants sits at the heart of academic work and is the cornerstone of modern scientific enterprise [Bohannon, 2013]. In the area of computer science, where most papers are submitted to large conferences, the fate of a paper is very much in the hands of automated assignment algorithms that help program chairs distribute thousands of papers among a similar number of committee members that serve as reviewers [Meir et al., 2021]. For this matching to happen, the committee members must first submit their preferences over papers, which are supposed to reflect both their competence and their interest in reviewing those particular papers, using a designated platform—a process typically referred to as *bidding*, and that many of the readers probably know well from their own experience. Cabanac and Preuss [2013] provide a detailed account of conference bidding and review flow. After the bidding process, one of the many algorithms for matching under preferences [Manlove, 2013; Klaus et al., 2016] can be used to find an assignment satisfying various notions of optimality, fairness, stability or other criteria [Chen et al., 2020; Aziz et al., 2015, 2019; Benabbou et al., 2021].

Crucially, the current design of the bidding process is far from eliciting the full preferences and capabilities of reviewers. First, in some widely used platforms there are only three levels of preference, typically ‘no’ / ‘maybe’ / ‘yes’. Other platforms including HotCRP are providing a finer scale for reviewers to express their preferences. However, it is not clear to what extent reviewers would use this flexibility, as *extreme responding* or *scale end bias* is a well known phenomena in many social sciences [Furnham, 1986]. E.g. YouTube removed the star rating scale since over 90% of responses were either one or five stars.¹ Additionally, it is not clear

¹<https://blog.youtube/news-and-events/five-stars-dominate-ratings/>

yet whether or not a finer grained scale of responses would actually lead to more desirable matchings between reviewers and papers. Second and more importantly, going over the entire list of submissions to determine the fit of every paper would take hours, whereas most reviewers would not invest that much time in bidding. Given that modern computer science conferences may have thousands of papers submitted to them, automated systems are being increasingly used to impute the bids of reviewers over papers, the best example being the Toronto Paper Matching System [Charlin and Zemel, 2013].

Hence, for these reasons and many others, it has been claimed that *skewed bidding*, i.e., where a few papers get many bids and some papers get no bids, is one of the main reasons for poor paper assignment [Fiez et al., 2020; Meir et al., 2021; Shah, 2021; Shah and Lipton, 2020; Lian et al., 2018]. This because some papers get insufficient (or no) bids and have to be assigned randomly or manually by the program chair, often ending up at unqualified reviewers.

For example, Cabanac and Preuss [2013] analyzed data from nearly 20,000 reviews in dozens of conferences managed on ConfMaster, and showed that more than 8,000 (42%) were done by reviewers who did not bid on the paper at all!²

Poor assignment, in turn, may affect both review quality [Stelmakh et al., 2019; Peng et al., 2017; Rodriguez et al., 2007]; and is also likely to put obstacles in the way of achieving alternative goals such as fairness [Payan and Zick, 2021; Lian et al., 2018], as creating a fair assignment crucially depends on actually knowing the preferences of the reviewers.

We note that simply increasing the bidding requirement, which increases the burden on reviewers during the bidding process, may still not be sufficient to deal with the issue of orphan papers. For example, at IJCAI 2018 each paper received almost 40 bids *on average*, and yet 140 papers (4%) had only two or fewer bids [Meir et al., 2021]. Given this, there have been two recent suggestions to alleviate the problem of skewed bidding in the literature:

1. present low-demand papers earlier (higher) on the list [Cabanac and Preuss, 2013; Fiez et al., 2020].
2. providing reviewers with information regarding paper demand [Meir et al., 2021];

Interestingly, the first suggestion builds on reviewers’ cognitive biases, whereas the latter one is exploiting their rational behavior.

In more detail, Fiez et al. [2020] proposed an algorithm to determine the order in which papers are presented to the reviewer during bidding, taking advantage of the ordering of papers to bidders. This suggestion rests on the *primacy effect*: items that appear earlier on a list are more likely to be selected [Murphy et al., 2006]. Primacy effects have been empirically shown to occur in conference bidding data on ConfMaster [Cabanac and Preuss, 2013]. The underlying idea is that demand can be smoothed by taking advantage of well known cognitive biases.

The other suggestion, by Meir et al. [2021], considered a model where the demand over papers is known (or revealed) to the bidders. They showed that as long as reviewers are individually rational and interpret their probability of being assigned a paper as inversely proportional to demand, a simple market-based scheme reduces the skew in bids and leads to an improved assignment. Drawing inspiration from the Trading Post Mechanism [Shapley and Shubik, 1977], they suggested tagging papers with their *inverse price* rather than the actual demand, and assign a *budget* that the bidder is encouraged to consume.

In both the work of Meir et al. [2021] and Fiez et al. [2020], the actual behavior of the individual bidder (i.e. how their bid is affected by primacy or by the demand) is *assumed*, and the theoretical and empirical results are contingent on these assumptions. However, bidding behavior with prices has never been tried or empirically validated, and while primacy effect has been shown to exist on average, it is not well understood how substantial it is compared to other factors.

²Indeed, ConfMaster also allows reviewers to express negative preference on a paper by bidding ‘no’, but this is not very helpful when facing thousands of papers.

Contribution Our main goal is to find out how different components of the bidding platform affect the probability that a participant will select a particular paper. The main motivation, following [Fiez et al., 2020; Meir et al., 2021] is to promote the selection of papers with few bids, thereby reducing the skew and indirectly improving the paper assignment. Yet the chairs may also have other reasons to promote some papers over others, and our research may be relevant for these as well.

Since previous work has suggested to control either the order of papers [Cabanac and Preuss, 2013; Fiez et al., 2020], or the information given to users on the demand [Meir et al., 2021], these are the main parameters we considered.

Hypothesis 1 (Order Effect) Subjects tend to select papers appearing earlier on the list.

Hypothesis 2 (Demand Effect) Subjects tend to select papers that are indicated as low-demand.

In addition we are interested in how these tendencies, if they exist, are distributed in the population, as well as in various factors affecting them. Hence, we designed and executed two types of experiments. The first is a field experiment on a medium-size workshop, and the second is a large scale experiment on Amazon Mechanical Turk where we control all the variables. In both experiments only some of the subjects were exposed to information on the demand, so their behavior can be compared to the control group.

Our main findings support both hypotheses, as we show that both paper order and information on demand can be used to shift reviewers towards low-demand papers.

However at the individual level there is a substantial difference: The order of papers has an effects on most subjects, but in a rather weak and inconsistent manner. In contrast, we identify in both experiments a small group of people that are *highly sensitive* to the demand, and results from the field experiment suggest that their effect on the bid distribution is substantial.

We further study via controlled experiments the relative and cumulative effect of exposing the subjects to different forms of information on the demand, and simple factors affecting compliance with the bidding instructions.

We conclude with a list of simple, practical suggestions to improve the use bidding platforms so as to reduce the prevalent skew in paper bidding, thereby improving paper matching.

1.1 Related Work

Ordering effects are well studied in economic and psychological models of choice. Typically, decision makers attend to the first few and last few items in a list more than the rest, increasing response rates for these items, see [Krosnick and Alwin, 1987] and references therein. In an academic context, papers appearing earlier on an email digest are more likely to be downloaded and cited [Feenberg et al., 2017]. Cabanac and Preuss were the first to show that ordering effects occur in paper bidding. Later, Fiez et al. [Fiez et al., 2020] suggested a sophisticated sorting algorithm that takes into account both dynamic demand and estimated reviewers’ preferences.

Rodriguez et al. [2007] aimed at uncovering the factors underlying bidders’ behavior in the JCDL’05 conference. Their starting point was that bids are expected to reflect the (objective) expertise of the reviewer w.r.t. the domain of the submission. They evaluate this expertise through alternative means, e.g., co-author network, keyword occurrence. The authors find very low correlation between reviewers’ areas and their bids, and conjecture that *reviewer fatigue* may be responsible for sloppy bidding. Our work does not get into whether reviewers’ preferences are indeed based on expertise (as opposed to, say, curiosity and personal interest in the title). It does however shed light on the other, more consistent factors that affect bidding behavior.

The figure consists of two side-by-side screenshots of a web-based bidding interface.

Left Screenshot: This is the standard bidding interface. At the top, there are buttons for 'Yes', 'Maybe', and 'Save', along with a 'Submit and finish' button. Below these, a summary shows 'yes bids: 3', 'maybe bids: 1', and 'Total bids: 4/40'. A search bar labeled 'Filter by word(s):' is present, with a 'filter' button and a 'Cancel filter/sort' button. The main table has columns: 'Choice', 'Paper ID', 'Title', 'Abstract', and 'Keywords'. The 'Choice' column has radio buttons for 'yes', 'maybe', and 'no'. The 'Abstract' column has a 'Show abstract' link. The 'Keywords' column contains text like 'recommender systems, matrix factorization, auc, local auc,'.

Right Screenshot: This is an enhanced version of the interface. It includes a 'Bidding requirement' section at the top right showing '1000'. Below this, a summary shows 'yes bids: 6', 'maybe bids: 1', and 'Total points: 310/1000'. A red box highlights the 'yes bidding points: 260p', 'maybe bidding points: 50p', and 'Total points: 310/1000'. The search bar is labeled 'algorithm'. The main table has an additional column, 'Bidding points', which is circled in red. The 'Abstract' column now has a 'Show abstract' link in red text.

Figure 1: Left: Example of our bidding interface used for all experiments. Right: with the additional column for iPrices (called ‘bidding points’) and the budget.

Assignment Algorithms The assignment of papers to reviewers is formally a version of the multi-agent resource allocation problem with capacities [Bouveret et al., 2016] and has been well studied in a number of areas of computer science [Goldsmith and Sloan, 2007; Lian et al., 2018], economics [Budish and Cantillon, 2012], and beyond [Dickerson et al., 2014]. Garg et al. [2010a] provide a comprehensive discussion of assignment algorithms, their application to the review process, and different methods for evaluating the quality of an assignment from both the conference and reviewer standpoint. Two popular ways to evaluate assignments are maximizing either the egalitarian welfare [Demko and Hill, 1988], i.e., making sure the worst off reviewers is as happy as can be or the utilitarian welfare, i.e., maximizing the sum of reported utilities for assigned papers across all reviewers. There are other refinements of these solution concepts [Garg et al., 2010a; Lian et al., 2018], and a large literature on calibrating feedback across reviewers for better assignment [Wang and Shah, 2019]. While the workshop in which we ran our field experiment used the utilitarian maximal assignment (an assignment maximizing social welfare), the results we report are independent of the assignment algorithm in use.

2 Experimental Design

We implemented a platform that resembles common paper bidding platforms—mainly EasyChair and ConfMaster.³ An example of the interface is shown in Figure 1.

2.1 The Basic Platform

In all experiments, the participant is presented with a table containing all possible papers. For each paper, the table specifies the title and keywords. In addition, the user may click a paper to expand and read the abstract. The user can bid on each paper using a radio button whose states are No/Maybe/Yes, where No is the default option. As is common in bidding platforms, we implemented basic search and filtering capabilities. The user may type a string in order to see only the papers containing this exact string anywhere in the title, keywords, or abstract. At the top, the user also sees how many papers have been marked as Yes and as Maybe so far, and may alter her selection of the papers at any time. Note that subjects could sort papers according to any column. The initial order depends on the experiment condition.

In some conditions additional information or options were provided in the interface including the inverse price of papers (called ‘iPrice’ in [Meir et al., 2021] and here, and ‘bidding points’ on the platform) or the total bidding requirement, marked in Fig. 1(Right). We discuss each of these design modifications in their respective experimental section.

³See <https://easychair.org/> and <https://confmaster.net/>.

iPrices and demand Following Shapley and Shubik [1977]; Meir et al. [2021], we define the *inverse Price* (iPrice) of a paper j as $p_j := 100 \cdot \min\{1, \frac{r}{d_j}\}$, where d_j is the current number of bids on the paper, and r is the number of copies of the paper that need to be assigned (throughout this paper $r = 3$). Thus a high iPrice indicates current low demand for the paper.

2.2 Field Experiment

For our field experiment, we used the bidding phase of the COMSOC-2021 international workshop.⁴ We partitioned the set of 42 reviewers randomly into a **field treatment (FT) group** consisting of 28 people that saw papers' *iPrices* during bidding, and a smaller **field control (FC) group** of 14 people that saw *no iPrices*. There were 93 submissions in total.

Bidding Process Both groups used our platform for bidding, where all 93 submissions were available along with the search and bidding interface shown in Figure 1. Using this interface, reviewers could also use the platform to report a conflict of interest on papers, but this was scarcely used. The control group had no extra information on demand and were asked to bid positively on at least 12 papers, of which 5-7 will be assigned. The treatment group saw the iPrices as in Fig. 1(Right), and had a budget of 800 bidding points. These bidding minimums for both groups were purely instructive and were not actively enforced in any way: reviewers could bid on any number of papers. The iPrices were set as explained above and updated on every new login, hence iPrices were static during a session but may change *between* sessions if an individual reviewer logged back in. We implemented the two caveats recommended in [Meir et al., 2021]: (a) the current bidder is always counted as a positive bid on all papers, to prevent price change during the bid; and (b) demands were initialized as uniform rather than empty to prevent a *cold start*. In practice only three reviewers logged in more than once to update their bids. Papers in both groups were initially presented according to their order of submission.⁵

Assignment The workshop used the bids entered by the committee members as input to a standard utilitarian maximal assignment algorithm with demands and conflict of interests [Lian et al., 2018; Garg et al., 2010b]. The implementation was the same as that of Lian et al. [2018] which uses Gurobi to solve the assignment ILP and allows for a range of paper and reviewer capacities, each reviewer was assigned 6 or 7 papers. The utility of the overall assignment used, using the bids as a proxy for reviewer utility, was 520.0. While there may be multiple assignments with the same utility we took the first one that Gurobi provided.

2.3 Controlled Experiments

In the controlled experiment we had a **Base (B) group** (same interface as the control group in the field experiment), and several different treatment groups. The main treatments we used were: revealing papers' iPrices to subjects in the **Price (P) group**; and visually highlighting low-demand papers in the **Highlight (H) group**. Additional conditions designed to study specific questions will be explained below.

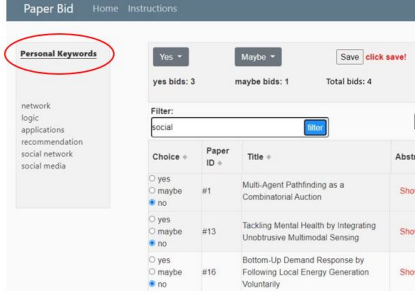
All subjects faced the same set of 550 papers from AAAI'15, who are publicly available.⁶ Subjects in the base group were requested to bid positively on 30 or on 40 papers, of which 8 will be assigned.

Setting Paper Demand Since subjects are participating independently of one another, we needed to generate the demand (i.e. the iPrices) for each paper. See Appendix A for details.

⁴<https://comsoc2021.net.technion.ac.il/>

⁵In hindsight it would have been better to present them in random order, as in the controlled experiments.

⁶<http://www.aaai.org/Library/AAAI/aaai15contents.php>.



Paper index	iPrice	Title	Reward
1	30	Blue Robots	0
2 *	80	Red Algorithms	2
3	40	Green Programs	0
4 *	100	Yellow Networks	0
5	0	Black Graphs	1
6	40	Red Databases	1
7	70	White Cyber	0

Figure 2: Left: A screenshot of the controlled experiment. The bidding interface is the same. The personal keywords representing the subject’s interests are highlighted on the left. Right: Example of calculating sensitivity parameters. Selected papers are marked with *. The reward is shown for a subject whose personal keywords are {**Red, Graphs, Algorithms**}.

Assignment While the assignment in the controlled experiment plays no role in our analysis, we describe it in Appendix A for completeness. The assignment was displayed to the participant immediately after they submitted their bid, together with the breakdown of the reward.

Incentives In our controlled experiment, participants were not actually reviewing any paper and thus a-priori had no incentive to prefer one paper over another. To mimic the situation of a reviewer trying to select ‘relevant’ papers, we assigned to each participant a set of six ‘personal keywords’ that supposedly reflect her interests. Subjects earned ‘coins’ for each of the 8 papers that were eventually assigned, and how many of these personal keywords they contained (either in the title or in the paper keywords or in the abstract). Each coin increased the bonus by \$0.25, thereby creating an incentive to bid on relevant papers.

The personal keywords were selected at random for each participant from the pool of all papers’ keywords, with constraints to make sure all participants had a similar amount of relevant papers. These personal keywords were displayed on the left side of the screen, see Fig. 2.

All participants expressed informed consent, and could terminate their participation at any time. The study was approved by the IRB of the authors’ institution.

2.4 Measuring Behavior

Since bidding behavior can be complex and depend on many variables, we develop simple measures that we can compare across subjects and groups of subjects.

For a set of papers S , we denote by $C(S) \subseteq S$ the subset of papers that were selected by subjects. A paper presented to two different subjects is would be counted as two separate papers. Also Note that we treat any positive bid (‘Maybe’/‘Yes’) as a selection. In particular, C_i is the set of papers selected by subject i .

We denote by $\bar{C}(S) := S \setminus C(S)$ the set of papers from S that were not selected, similarly, $\bar{C}_i$ are the papers not selected by subject i .

We denote by $p_s \in [0, 100]$ the iPrice of paper s . In the field experiment, the iPrice was derived from the actual demand as explained above, and was updated with every new login; whereas in the controlled experiment it was generated once per subject and remained fixed.

Measuring Individual Behavior For each of the features we used $X \in \{R, O, D\}$ (for (R)eward or Relevance, (O)rder, and (D)emand, respectively) and each paper $s \in S_i$, we denote by $f^X(s) \in [0, 1]$ the relevant feature of the displayed paper.

In the example in Fig. 2(Right) paper #3 has $f^O(s) = \frac{3}{7}$, $f^D(s) = \frac{80}{100}$, and $f^R(s) = \frac{0}{2}$, as the maximal reward in this example is 2 (the reward scheme in the paper was more involved, see instructions).

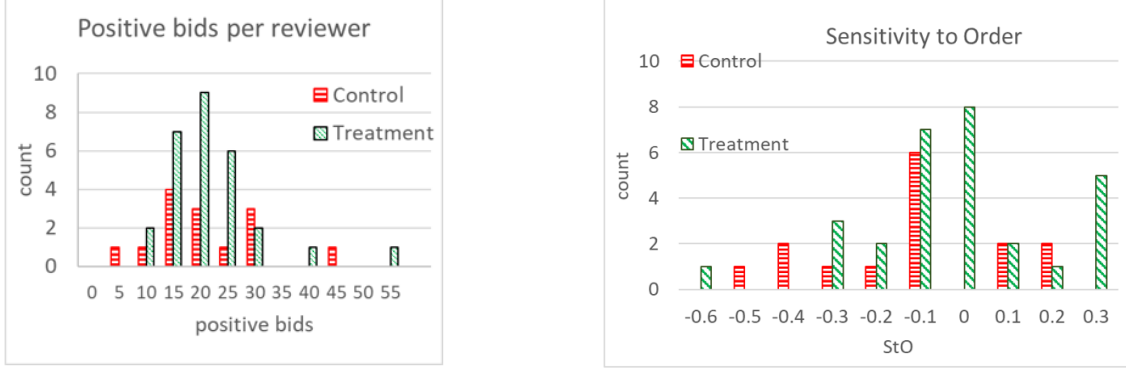


Figure 3: Left: Histogram of the number of bids for each reviewer in the field experiment. Right: Distribution of individual order sensitivity values in the field experiment.

For a subset of samples S' , we used the average: $f^X(S') := \frac{1}{|S'|} \sum_{s \in S'} f^X(s)$. E.g. for $S' = \{1, 2, 3\}$ in our example, we have $f^D(S') = \frac{1}{3}(0.3 + 0 + 0.8) \cong 0.366$.

For every subject $i \in N$ and feature $X \in \{R, O, D\}$, we defined the ‘sensitivity-to- X ’ as the difference between the average value of the feature in selected and unselected papers. Formally:

$$StX_i := f^X(C_i) - f^X(\bar{C}_i). \quad (1)$$

StX_i is always in $[-1, 1]$, and its expected value is 0 if i is completely insensitive to feature X (e.g. selects papers at random). For the subject in our example, where the selected papers are $C_i = \{2, 4\}$ and $\bar{C}_i = \{1, 3, 5, 6, 7\}$, we have

- $StR = 0.5 - 0.1 = 0.4$, indicating a moderate sensitivity to reward;
- $StO = \frac{6}{14} - \frac{22}{35} = -0.2$, meaning the subject tends to select earlier papers; and
- $StD = 0.9 - 0.36 = 0.54$, meaning sensitivity towards paper with low demand (=high iPrice).

Note that StR cannot be evaluated in the field experiment since we have no direct access to the reviewers’ real preferences and expertise.

Measuring Group Behavior One way to measure the group behavior is considering the average StX values of group members. When we want to condition on other attributes, we measure the probability of selecting a paper as a function of the relevant feature (e.g. initial position in the table), while controlling for relevance. Formally, given a set of samples S' (say, ‘all papers in the second quantile of positions that are highly relevant to their respective subject’), the probability of selection is $PS(S') := \frac{|C(S')|}{|S'|}$.

3 Results from the Field Experiment

3.1 Distribution of Bids

The empirical distribution of bids is shown in Figure 3(Left). In the control group there were a total of 267 bids, 19.1 bids per user, while for the treatment group there were 547 bids, which are 19.5 bids per user.

Sensitivity to Order According to Table 1, only the control group demonstrated sensitivity to order, however the effect is barely statistically significant (presumably due to the small number of reviewers on that group).

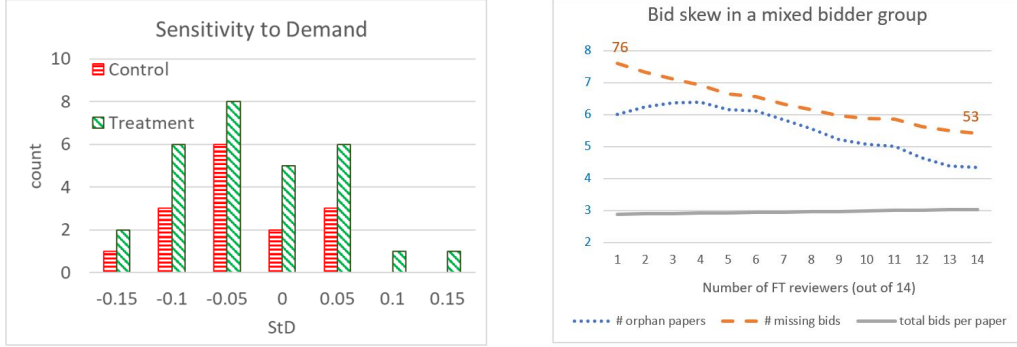


Figure 4: Left: Distribution of demand sensitivity values in the field experiment. Right: Bootstrap results for number of underdemanded and orphan papers in mixed group of 14 reviewers.

Sensitivity to Demand The first observation from Table 1 regarding demand sensitivity is that it is negative in both groups, on average. This may seem surprising but actually makes intuitive sense as in a real conference there is positive correlation in bids, i.e., you are more likely to bid on a popular paper. Hence, only the *difference* between the groups matters. The average of StD is slightly higher in the treatment group, but this is not statistically significant (see Appendix D). It is more instructive to look at the distribution of StD values (Fig. 4(Left)): we can see clearly that in the treatment group there are several subjects that are highly sensitive to high iPrices (i.e. to low demand), whereas the distribution of the others is similar to the control group.

3.2 Skewed Bids

We compared the number of papers that were under-demanded in each group, that is, received fewer than the 3 bids necessary to find a good assignment. In the control group there were 47 papers that received fewer than three bids, with 6 of these being papers that received no bids at all. For the treatment group there were only 6 papers that were under-demanded and only a single paper that received no bids. However, this must be partially due to the difference in the size of the two groups.

To address this we looked both at the number of orphan papers and at the number of *missing bids* (minimal additional bids required so that every paper has at least 3 bids) that would appear under a bootstrap sampling paradigm [Bruce et al., 2020]. To do this we took the set of bids and sampled a “small committee” from each group with 14 reviewers in it 1000 times. As we can see in Fig. 4(Right), although the average number of bids remains unchanged, the number of missing bids and orphans drops significantly as we replace FC bidders with FT bidders, indicating that even the small number of demand-sensitive bidders have a substantial effect on the bid skew.

3.3 Discussion of the Field Experiment

The initial results from our field experiment suggest that: (1) there seems to be a weak order effect; (2a) there is some fraction of reviewers that are highly sensitive to the demand when given via bidding points and budgets; (2b) this increased sensitivity to demand reduces the number of missing bids and orphan papers; (3) subjects who had budgets were more compliant, possibly due to differences in the UI.

However the small number of reviewers makes it difficult to make any strong conclusion. In addition some parameters cannot be controlled (such as inherent demand for papers); or were not controlled in our design (such as paper order or displaying the bidding requirement). We therefore turn to controlled experiments to better understand these effects.

Code	Condition	participants	non-spammers	games	StReward	StOrder	StDemand
B	Base	50	29	29	0.20 ± 0.06	-0.11 ± 0.10	$-0.01 \pm 0.03\#$
P	iPrices	124	80	80	0.21 ± 0.04	-0.16 ± 0.07	0.08 ± 0.04
H	Highlight	43	21	39	0.19 ± 0.08	-0.13 ± 0.07	0.05 ± 0.04
PS	P + Sort	34	17	17	0.22 ± 0.09	—	0.14 ± 0.10
PHS	P + H + Sort	33	28	59	0.24 ± 0.06	—	$0.09 \pm 0.10\#$
Total (controlled exp.)		284	175	224			
FC	Control	14	14	—	—	-0.12 ± 0.11	-0.04 ± 0.03
FT	Treatment	28	28	—	—	-0.03 ± 0.08	-0.03 ± 0.03

Table 1: The left side shows number of participants and played games in each group in the controlled experiment. The right columns show the average sensitivity of each group to each parameter, within 2 standard errors. We mark with # results in the controlled experiment that do not statistically differ from 0.

4 Controlled Experiments

Conditions Our **base group (B)** was similar to the control group at the field experiment, except that papers were displayed at a random order, and we added the bidding requirement to the UI in order to rule out this as a potential source of differences between groups. See Fig. 1.

In addition to the base group, we had the following treatments.

iPrices (P) In this condition (similarly to the FT group in the field experiment) subjects had an additional column titled ‘Bidding points’ showing papers’ iPrices as integers in the range $[0, 100]$. The bidding requirement was set as a ‘budget’ of 1000 points.

Highlight (H) In this condition we did not show the iPrice, but instead highlighted low-demand papers in green (when iPrice is 100) or yellow (when iPrice in $[70, 99]$, see Figure 5):

iPrices + Sort (PS) Similar to Condition P, except papers were initially sorted by increasing demand (decreasing iPrice).

iPrices + Highlight + Sort (PHS) Similar to PS, with also highlighting low-demand papers as in Condition H.

Data Collection We collected data from 284 participants on Amazon Mechanical Turk.⁷ Subjects were allowed to play up to three times. Participants were randomly assigned to the base group or to one of the treatment groups. The total number of participants of each group appears in the second column in Table 1. The threshold for rejection was set at 12 coins.

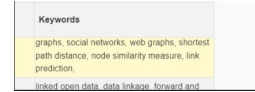


Figure 5: Highlighting.

Spammers and Sensitivity to Relevance There was a distinctive group of subjects who did not respond to paper relevance (‘spammers’) and were not included in the rest of the analysis. We explain this in detail in Appendix B.

To better understand the isolated effect of each factor, we start by analysing the Base condition and conditions (H)ighlight and i(P)rices. For Example, the StR column in Table 1 shows that in all groups the mean sensitivity (of non-spammers) is about 0.2, and is significantly higher than 0.

⁷In one of the early trials that were not included in the final study, we used undergraduate students at the authors’ local institution. Results were qualitatively the same.

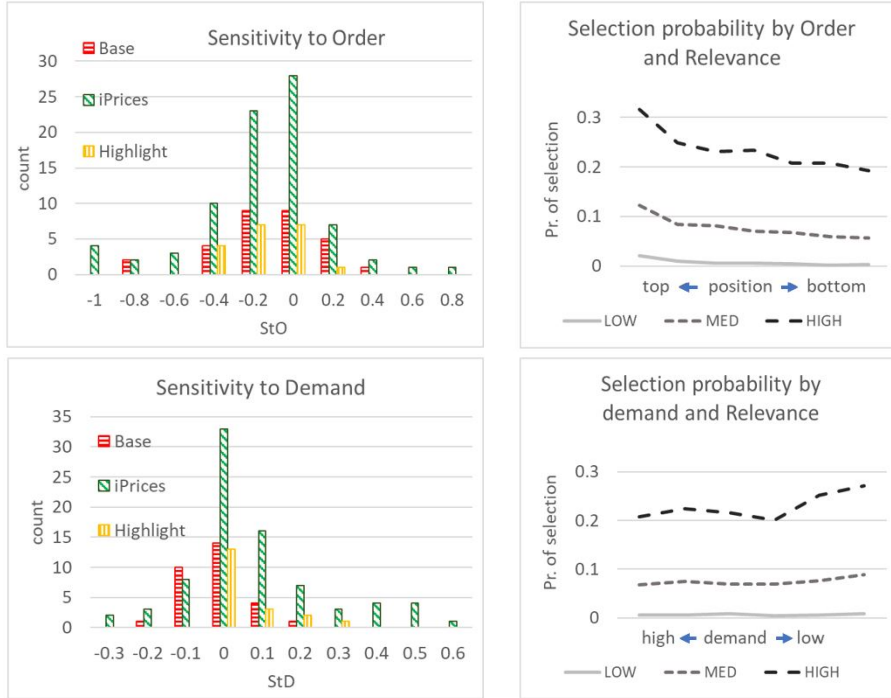


Figure 6: Left: Histogram of individual StO (top) and StD (bottom) values. Right: Selection probability of a paper, conditional on its position and relevance (top), and on its demand and relevance (bottom). Probability is calculated over all subjects in conditions B,P,H. LOW/MED/HIGH relevance means that the paper contained 0, 1, or more relevant words, respectively.

4.1 Paper Order

We can see that in all three conditions, there is similar average sensitivity to order, of about -0.13, i.e. there is a statistically significant bias to papers that appear earlier. However reward still plays a more important role in selection.⁸

But are all subjects slightly biased or is it a small number of highly biased subjects? For this, we look at the distribution of individual StO values in our controlled experiment (Fig. 6, top left).

From the figure, it seems that most subjects are prone to some bias (sensitivity is most often negative but not below -0.4); yet there is a non-negligible number of subjects with a very strong sensitivity, which essentially marked papers at the very top. Some subjects had high positive StO values, meaning they deliberately marked papers at the bottom of the list.

Another question we can ask is whether all papers are equally likely to be promoted when appearing earlier. As we can see in Fig. 6 (top right), primacy affects irrelevant and relevant papers alike, where selection probability drops sharply for papers that are not at the top, and then continues to decrease moderately.

Our findings regarding order effect are largely consistent with those of Cabanac and Preuss [2013] from real conferences, and thus support our Hypothesis 1. The added value of our controlled experiments is three-fold: how order effects are distributed across the population; the dependence (or lack of thereof) on the relevance of the paper; and individual consistency, which we consider next.

⁸For many spammer subjects, the StO was even more negative, which is not surprising or interesting.

Consistency A third question we may ask is whether the bias towards early papers is *consistent*. We analyzed the behavior of subjects who played two or three times, comparing their StO measure each time. We find that the gap between two games of the same subject is the same, on average, as between two independent subjects.

This, together with the findings above, suggests that order effects, while substantial, act mostly at an unpredictable and inconsistent manner.

4.2 Paper Demand

We considered two ways to communicate papers' demand to subjects. The first was adopting the market scheme of Meir et al. [2021] where low-demand papers have high iPrices (condition P). In condition H we simply highlighted the low-demand papers visually.

Sensitivity to Demand The right column in Table 1 shows that the Base group is completely insensitive to the demand (as expected, since they have no information about it); the iPrice scheme is moderately effective; and highlighting alone has a low effect (barely statistically significant). Looking at the distribution of sensitivity to demand in Fig. 6 (bottom left), we can see that in contrast to the primacy effect, most subjects in conditions P and H are not sensitive to the demand. The effect we see is due to a relatively small number of highly sensitive subjects. This corroborates our initial finding from the field experiment, and supports our Hypothesis 2.

Price Scheme More Effective than Highlighting We can see in Table 1 that the effect of highlighting papers by itself is borderline significant (also only 3 of the 21 subjects demonstrated significant bias towards highlighted papers in Fig. 6). In contrast, about third of the subjects who were exposed to iPrices were significantly affected, and the overall bias was double.

Which Papers are Affected? We can see that the effect of high iPrices is mainly on papers that are already relevant (Fig. 6, bottom right). This is another difference from the effect of paper order. It is also another evidence of rational decision making (in the economic sense), as the iPrice indicates the probability of getting the paper.

Consistency In contrast to order effects, sensitivity to demand is consistent and predictable: for subjects who played more than once, the gap between two games of a subject is less than half of the gap from other subjects.

4.3 Using All Treatments?

Since paper order, iPrice and highlighting all have some positive effect, it might make sense to combine them together in order to influence people to spread their bids even more.

In particular, we can adopt the suggestion of Cabanac and Preuss [2013] to actively sort papers by increasing demand *and* combine this with Meir et al. [2021] iPrice scheme. Further, we can highlight the most underdemanded papers in the table.

We therefore ran another experiment with two more groups: In group P+Sort we displayed iPrices and budget as in condition P *and* sorted the papers initially by decreasing iPrice (so underdemanded papers are on top); In group P+H+Sort we did the same *and* highlighted the underdemanded (high-iPrice) papers as in condition H. Note that in these two conditions StO is the same as (negative) StO so not shown.

By looking back at the StD results in Table 1, we can see that in group P+Sort there is more sensitivity to demand, as expected. It is also higher than the mean StO of subjects in the previous conditions, indicating some synergy between the price scheme and the initial order.

However highlighting papers in addition did not help at all (see condition P+H+Sort), and in fact only added noise. We analyse possible reasons for this strange phenomenon in Appendix E.

5 Discussion

Our combined experiments in bidding behavior show that:

1. Bidding likelihood increases uniformly for papers appearing higher in the list (corroborating previous empirical findings);
2. Presenting papers' demand in the form of iPrices positively influences a small but non-negligible subset of people to shift their selection to low-demand papers;
3. Presenting the bidding requirement during bidding (rather than just include it in the instructions beforehand) results in much higher compliance (see Appendix C).

Our field experiment further showed that shifting the demand of even few bidders towards low-demand papers, reduces the skew in bids and makes sure more papers get the minimal required amount of bids.

Critique There are two main concerns about the validity of our results. First, there is an internal validity issue: One can ask whether the behavior we see is consistent or sporadic. This is important as consistency also means predictability. Indeed at least with regards to ordering effects, our results showed that individuals were not consistent in their behavior, and thus there is risk in constructing mechanisms that assume a particular order-influenced behavior.

Another concern is external validity: will the behavior of researchers bidding on real papers be similar to that of AMT workers who play a game for recreation and/or money?⁹ While we cannot guarantee that this is the case, the similar patterns found in the field experiment suggest that this is a reasonable working assumption.

Practical Recommendations We believe that adopting the simple market scheme of Meir et al. [2021] can have a positive influence on distribution of bids during bidding phase. This influence can be increased by combining other UI factors such as highlighting and/or use the current demand as a factor in sorting presented papers [Cabanac and Preuss, 2013; Fiez et al., 2020]. Regardless of the bidding scheme, we recommend that the bidding requirement (in terms of number of positive bids or budget) will be displayed during bidding. These changes can be easily implemented in existing platforms such as EasyChair and ConfMaster, and be offered to conference organizers as optional features.

That said, we highly recommend to do these changes carefully:

- Consult UX/UI experts regarding the best way to highlight papers so as to avoid confusion, choosing the best term for the iPrices and budgets, etc.;
- Explain reviewers/committee members that they can bid as they wish (even ignore all additional information), but will be more likely to get their desired papers by following the bidding instructions;
- As for paper order, we should keep in mind that most platforms offer the user flexibility in how to sort the papers, so users should have to option to choose whether demand should be a factor in this order;
- Test suggested changes on a subset of conference participant and/or in smaller workshops before full adoption.

We hope these suggestions will contribute to improving the review process for all.

⁹The analysis in Appendix E raises one such concern.

References

- Haris Aziz, Serge Gaspers, Simon Mackenzie, and Toby Walsh. 2015. Fair assignment of indivisible objects under ordinal preferences. *Artificial Intelligence* 227 (2015), 71–92.
- Haris Aziz, Xin Huang, Nicholas Mattei, and Erel Segal-Halevi. 2019. The Constrained Round Robin Algorithm for Fair and Efficient Allocation. *arXiv preprint arXiv:1908.00161* (2019).
- Nawal Benabbou, Mithun Chakraborty, Ayumi Igarashi, and Yair Zick. 2021. Finding fair and efficient allocations for matroid rank valuations. *ACM Transactions on Economics and Computation* 9, 4 (2021), 1–41.
- J Bohannon. 2013. Who’s Afraid of Peer Review? *Science* 342, 6154 (2013), 60–65. <https://doi.org/10.1126/science.342.6154.60>
- S. Bouveret, Y. Chevaleyre, and J. Lang. 2016. Fair Allocation of Indivisible Goods. In *Handbook of Computational Social Choice*, F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia (Eds.). Cambridge University Press, Chapter 12, 284–311.
- Peter Bruce, Andrew Bruce, and Peter Gedeck. 2020. *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O’Reilly Media.
- E. Budish and E. Cantillon. 2012. The Multi-unit Assignment Problem: Theory and Evidence from Course Allocation at Harvard. *The American Economic Review* 102, 5 (2012), 2237–2271.
- Guillaume Cabanac and Thomas Preuss. 2013. Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. *Journal of the American Society for Information Science and Technology* 64, 2 (2013), 405–415.
- Laurent Charlin and Richard Zemel. 2013. The Toronto paper matching system: an automated paper-reviewer assignment system. (2013).
- Jiehua Chen, Robert Ganian, and Thekla Hamm. 2020. Stable matchings with diversity constraints: Affirmative action is beyond NP. *arXiv preprint arXiv:2001.10087* (2020).
- S. Demko and T. P. Hill. 1988. Equitable distribution of indivisible objects. *Mathematical Social Sciences* 16 (1988), 145–158.
- J. P. Dickerson, A. D. Procaccia, and T. Sandholm. 2014. Price of fairness in kidney exchange. In *Proceedings of the 13th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 1013–1020.
- Daniel Feenberg, Ina Ganguli, Patrick Gaule, and Jonathan Gruber. 2017. It’s good to be first: Order bias in reading and citing NBER working papers. *Review of Economics and Statistics* 99, 1 (2017), 32–39.
- Tanner Fiez, Nihar Shah, and Lillian Ratliff. 2020. A SUPER* algorithm to optimize paper bidding in peer review. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 580–589.
- Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personality and individual differences* 7, 3 (1986), 385–400.
- N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre. 2010a. Assigning Papers to Referees. *Algorithmica* 58, 1 (2010), 119–136.

- Naveen Garg, Telikepalli Kavitha, Amit Kumar, Kurt Mehlhorn, and Julián Mestre. 2010b. Assigning papers to referees. *Algorithmica* 58, 1 (2010), 119–136.
- J. Goldsmith and R. Sloan. 2007. The AI onference paper assignment problem. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI) Workshop on Preference Handling for Artificial Intelligence (MPREF)*.
- John L Hodges. 1958. The significance probability of the Smirnov two-sample test. *Arkiv för Matematik* 3, 5 (1958), 469–486.
- Bettina Klaus, David F. Manlove, and Francesca Rossi. 2016. Matching under Preferences. In *Handbook of Computational Social Choice*, Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (Eds.). Cambridge University Press, 333–355. <https://doi.org/10.1017/CB09781107446984.015>
- Jon A Krosnick and Duane F Alwin. 1987. An evaluation of a cognitive theory of response-order effects in survey measurement. *Public opinion quarterly* 51, 2 (1987), 201–219.
- Jing Wu Lian, Nicholas Mattei, Renee Noble, and Toby Walsh. 2018. The Conference Paper Assignment Problem: Using Order Weighted Averages to Assign Indivisible Goods. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI’18)*. 1138–1145.
- David Manlove. 2013. *Algorithmics of matching under preferences*. Vol. 2. World Scientific.
- Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.
- Reshef Meir, Jérôme Lang, Julien Lesca, Nicholas Mattei, and Natan Kaminsky. 2021. A Market-Inspired Bidding Scheme for Peer Review Paper Assignment. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*. 4776–4784.
- Jamie Murphy, Charles Hofacker, and Richard Mizerski. 2006. Primacy and recency effects on clicking behavior. *Journal of computer-mediated communication* 11, 2 (2006), 522–535.
- Justin Payan and Yair Zick. 2021. I Will Have Order! Optimizing Orders for Fair Reviewer Assignment. *arXiv preprint arXiv:2108.02126* (2021).
- Hongwei Peng, Haojie Hu, Keqiang Wang, and Xiaoling Wang. 2017. Time-aware and topic-based reviewer assignment. In *International Conference on Database Systems for Advanced Applications*. Springer, 145–157.
- Marko A. Rodriguez, Johan Bollen, and Herbert Van de Sompel. 2007. Mapping the bid behavior of conference referees. *J. Informetrics* 1, 1 (2007), 68–82.
- Nihar B Shah. 2021. Systemic Challenges and Solutions on Bias and Unfairness in Peer Review. *Working Paper* (2021).
- Nihar B Shah and Zachary Lipton. 2020. TheWebConf 2020 Tutorial on Fairness and Bias in Peer Review and other Sociotechnical Intelligent Systems (Part II on Peer Review). Tutorial notes.
- L. Shapley and M. Shubik. 1977. Trade using one commodity as a means of payment. *Journal of Political Economy* 5, 85 (1977), 937–968.
- Ivan Stelmakh, Nihar B Shah, and Aarti Singh. 2019. PeerReview4All: Fair and accurate reviewer assignment in peer review. In *Algorithmic Learning Theory*. PMLR, 828–856.

Jingyan Wang and Nihar B Shah. 2019. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 864–872.

A Further Details on the Controlled Experiment Setup

Setting papers’ demand In practice, rather than generating artificial demand, we sampled the iPrice directly from a uniform distribution on $[-25, 120]$, and truncated to the range $[0, 100]$. This is to guarantee we cover the entire range and also have a substantial number of papers with extreme iPrices. Although in reality no paper could have an iPrice of 0 (as it indicates infinite demand), we still wanted to see how this will affect behavior.

Assignment algorithm After selection, we determined the assignment based on subject’s bids using a simple randomized allocation: The iPrice of each paper (see Section 2.4) is used as an approximate assignment probability. The algorithm first considers all ‘Yes’ bids in random order and assigns papers with probability p_j until 8 papers are assigned or until all of ‘Yes’ bids have been considered. It then repeats the process for ‘Maybe’ bids and then for all other papers.

Participants were not aware of the exact allocation algorithm, but were told that papers with positive bids were more likely to be assigned, and that the chance also depends on the demand for the paper (to which they may or may not be exposed according to the condition they are in).

Instructions and Demo To make sure that the (rather complex) instructions of our experiment are understood, we used: detailed instructions; an online quiz (see appendix); and a demo game. We also informed participants up front that failure to reach minimal required reward may result in rejection of the job—standard for conducting online behavioral research and adapted to our problem [Mason and Suri, 2012].

The instructions and quiz focused on explaining that the payment depends only on the assigned papers (8 in total) and not directly on the bid. The demo was very similar to the game except it only contained 50 papers and 3 personal keywords. Participants that did not reach the minimal required reward in the demo could not continue to the game but could try the demo again (up to 3 times in total).

B Spammers and Sensitivity to relevance

The first thing we had to check was whether subjects are responsive to incentives at all, as otherwise there is not much value in measuring their behavior.

We had an internal criterion used to filter out participants that had consistently failed to select relevant papers. For this, we calculated for each subject her sensitivity to relevance (StR).

Note that a completely random selection should result with StR that is close to 0, whereas even a very weak tendency to select relevant papers (e.g. avoiding papers that have no personal keywords) should result in a strictly positive StR.

The empirical distribution of StR is visible in Fig. 7. We can see that there is a bimodal distribution, indicating two types of subjects: a large minority of subjects whose StR is distributed around 0 (left of the dashed line in the figure), and the majority of subjects, whose StR is well separated from 0.

We treat the first type as ‘spammers’, and do not include them in our behavioral analyses.¹⁰ The right column in Table 1 shows the number of usable subjects from each group. Note that our criterion for spammers is conservative, and that there may still be some spammers within the low-StR subjects we analyze.

¹⁰A participant may be a ‘spammer’ either if she plays without effort for a quick gain, or if she failed to understand and follow the reward structure. Some evidence suggest that most spammers are of the first type, as they had spent much less time on bidding, and provided unreasonable demographic information.

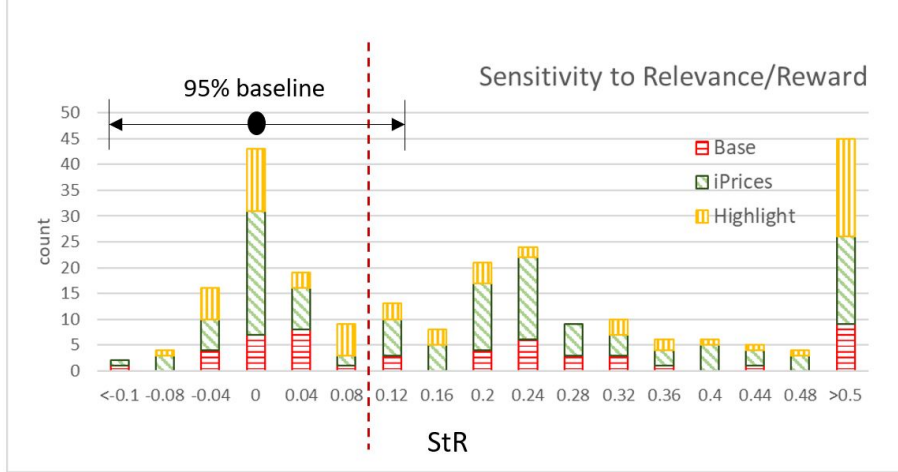


Figure 7: A histogram of subjects’ StR. Subjects of different conditions are separated by color and pattern. The arrow indicates the range in which we expect to find subjects who are completely insensitive. This is the 95% confidence interval for the StO value of a subject that picks at least 20 papers completely at random. Subjects who pick fewer papers have a higher chance of exceeding the interval.

Payment Rejection We used a more conservative criterion to reject AMT subjects.

We only rejected subjects who failed the announced criterion (at least 12 coins) *and* our internal ‘spammer’ criterion explained above. Thus spammers who (by chance or due to false classification) reached 12 coins were paid, but their data was not included in the analysis.

C Compliance with Bidding Instructions

A very visible difference between the groups at the field experiment is that in the treatment group the total bid amounts of each reviewer are *concentrated* (with a few outliers), whereas in the control group the distribution is scattered. This is even though the control group were instructed to bid on a given number of papers (12), and the instructions to the treatment group was in terms of bidding points. To more accurately measure this difference we define a new measure.

The *compliance ratio* of a subject with instructions to bid on R papers is computed as $Comp_i := \frac{|C_i|}{R}$. For subjects that have prices and budgets we compute the compliance as $Comp_i := \frac{\sum_{s' \in C_i} p_s}{B}$ where B is the budget.

In either case, a number below 1 indicates underbidding and a greater number indicates overbidding.

Indeed, for condition FC (control) the average compliance ratio was 1.59 (median 1.55) with a standard deviation of 0.75; whereas in condition FT most subjects were highly compliant, with average, median, and standard deviation of 1.15, 0.97, and 0.52 respectively.

Is the introduction of bidding points and budget responsible for this? By closely comparing the UI of both conditions, we see another possible explanation. While subjects in the treatment group can easily track their progress through the budget that is shown throughout bidding (see Fig. 1(bottom)), the interface of the control group followed the standard practice in bidding platforms where the total bids are displayed but not the bidding requirement. This requirement (12 papers in our case) is usually informally specified in an email to the reviewers, as it was in our field experiment.

We therefore formulate a third hypothesis:

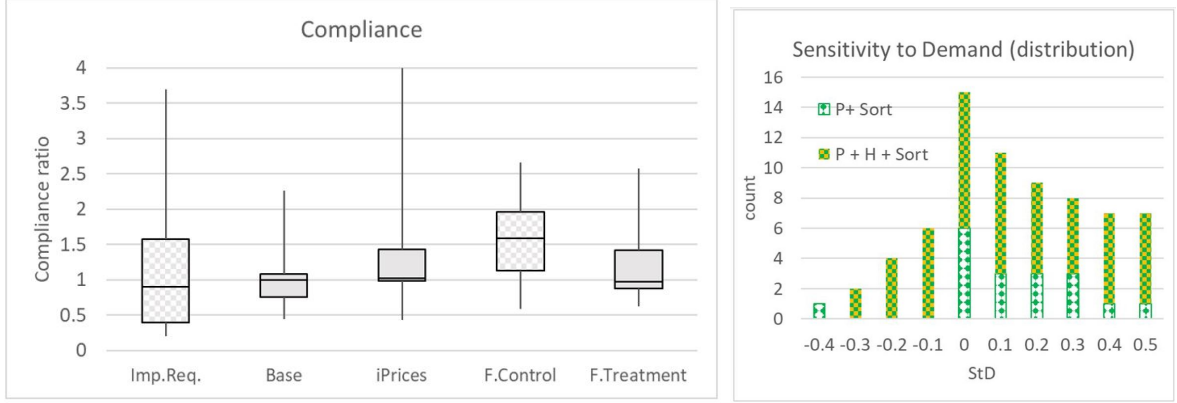


Figure 8: Left: A box plot of the compliance ratio distribution under different conditions (showing quartiles 1-3 and top and bottom 5 percentiles). In the checkered conditions (IR,FC) the bidding requirement was not shown during bidding.

Right: Distribution of individual StD values in the two new conditions.

Hypothesis 3 : Displaying the bidding requirement (regardless of other information) makes reviewers more compliant.

Note that by ‘more compliant’ we mean the compliance ratio of more subjects is closer to 1, not necessarily higher.

Controlled experiment In addition to the five conditions of the controlled experiment described in Sec. 4, we added a 6th group:

Implicit Request (IR) This condition was identical to the base condition, except that the bidding requirement did not appear on the screen during bidding (as in most conferences and in our FC group). It still appeared on the instructions, that were accessible throughout the experiment.

This group of subjects contained 36 non-spammers (out of 54 subjects in total), who played one game each.

C.1 Results

Comparing groups IR and B allows us to directly test Hypothesis 3. Indeed, we can see in the two leftmost columns of Fig. 8 that merely presenting the bidding requirement on the screen substantially boosts compliance: while the average compliance in both is close to 1 (1.14 and 1.07, respectively), the variance of IR is much higher (p -value is 0.001 in an F-test). This finding supports the hypothesis. Recall that subjects could view the instructions at any time even in Condition B, so the strength of the effect is rather surprising.

There was no substantial difference between groups P and B in terms of concentration, although there was less underbidding and more overbidding in group P. This is in contrast to what we saw at the field experiment, but recall that groups FC was more similar to IR than to B in this respect.

D Bid Distribution in Field Experiment

Looking at the StD statistics, the control group has a StD of -0.0432 with a standard deviation of 0.056, while the treatment group is -0.0259 with a standard deviation of 0.072. So we see that the group exposed to prices is more sensitive to them. Finally we wanted to see if the

induced bid distributions are drawn from different distributions. To test this we used a two sample Kolmogorov-Smirnov test with the null hypothesis that the treatment distribution was less than the control distribution [Hodges, 1958]. This resulted in a test statistic of 0.001429 and a p -value of 0.67, so we cannot reject the null hypothesis. So while we cannot say that bidders exposed to price provided significantly fewer bids, there is evidence to suggest this and we may be able to conclude this given a larger experiment.

E Strange Behavior in the P+H+Sort Condition

Recall that the mean sensitivity to demand in Condition PHS was substantially lower than in PS alone, even though the conditions are very similar, and Highlighting should only increase the sensitivity.

We can see a possible reason for this in the distribution of personal StD values in Fig. 8(Right): while almost all subject in group P+H+Sort were influenced, it seems that some of them deliberately selected *high-demand* papers! Moreover, for those subjects who played more than one game, this behavior was consistent.

While it is possible that our instructions in this condition only were not clear and led to confusion, there is also an alternative explanation.

An ‘exact bidder’ is a subject who bids exactly the budget. While this behavior was rare in general (occurred in about 10% of the games in the other conditions), in the PHS condition more than half the games were exact. Moreover, there were 11 (out of 28) users who were exact in all of their games.

While exact bidding is not a problem on its own, such an excessive rate of two rare behaviors (exact bidding and negative StD) in a single condition may indicate that there were few persons (or perhaps even one) behind many of the accounts.

While we cannot verify this conjecture, it is a reminder about how careful we should be when designing and analyzing online experiments, and of the need for independent corroborations.

When ignoring games with suspicious exact bidding, we get mean StD rate of 0.13 ± 0.09 and 0.15 ± 0.09 in the PS and PHS conditions, respectively. This indicates that the PHS condition is indeed the most effective in making bidders demand-sensitive and reduce bid skew.

F Experiment Instructions and Quiz

Instructions to participate in the bidding experiment

In this experiment, you are simulating a reviewer that needs to bid on papers to review.

You have a list of personal keywords that reflect your professional interests. Your reward will be determined according to the 8 papers you are assigned and how well they match your personal keywords. The experiment only includes the bidding part. **You will not need to actually read or review any paper.**

You will be able to play up to 4 times to earn more rewards.

Each time you will get a new set of personal keywords, a new assignment will be computed based on your bids. The new reward will be added to your payment. You may quit after fewer rounds to receive the rewards you earned so far.

Bidding

To bid on a paper, click on "Maybe" or "Yes".

You may click on "show abstract" to see the paper abstract.

You may also filter the list for your keywords or any other word.

We request that you bid positively (yes/maybe) on at least 1000 bidding points, but you do not have to follow this instruction and it will not affect your reward.

Important: you have to click "save" before moving to another page or tab (including the Instructions tab).

Assignment

When the experiment is over, you will be assigned **exactly 8 papers**.

The papers you will be assigned depend on your bid and on the demand for each paper (see "bidding points" below). That is, you are more likely to get a paper that you bid on. A "Yes" bid increases the chance more than a "Maybe" bid. You are also less likely to get highly demanded papers.

Reward

1 coin = \$0.25

In each round You will get a reward of 10 coins for your participation, plus a bonus that depends on your assignment.

Each of these 8 assigned papers will affect your reward according to its relevance:

- If the paper contains (in the **title, keywords, or abstract**) at least 2 of your personal keywords, then it is relevant for you, and worth 2 coins.

- If it contains at least 3 words, then it is highly relevant, and worth 6 coins.

- If it contains no relevant words at all, it is irrelevant, and we will deduct 1 coins from your reward.

IMPORTANT: while it is possible to earn up to 232 coins by playing carefully, **your entire job may be rejected if your average reward per round is less than**

15 coins.

It is therefore crucial to understand the rules and the reward structure.

Bidding points

Every paper is worth a certain amount of bidding points. A high number means **a higher chance to get the paper** due to lower demand.

You will see on the screen the sum of your current bidding points (over all papers with positive bids).

but you do not have to follow this instruction and it will not affect your reward.

Quiz and Demo

After reading the example below, you will take a simple quiz that will test your understanding of the instructions.

Then you will play a simplified demo to verify you understand the reward structure.

Failing to reach 15 coins will disqualify you from participating in the experiment (you will have two attempts).

Note, the demo only simulates the reward. Your actual reward will be set only according to the papers you are assigned in the real experiment.

Example

Jane has the personal keywords {**NLP, vectors, multiprocessing, LDA, quantum computing, coil**}.

Jane bids "Maybe" on papers {#34, #66, #193, #201}

Jane bids "Yes" on papers {#12, #25, #174, #196, #304, #332, #347, #454, #466}

After the experiment is over, Jane is assigned 8 papers:

{#174, #196, #304, #332, #347, #66, #3, #189}

#Paper assigned	Personal keywords	relevance	Reward (coins)
#174	LDA, vectors, multiprocessing	Highly relevant	+6
#196	LDA, coil	Relevant	+2
#66	Vectors		
#3	-	Irrelevant	-1
#304	Coil, quantum computing	Relevant	+2
#332	NLP, vectors, coil	Highly relevant	+6
#347	NLP, multiprocessing	Relevant	+2
#189	LDA		
Base reward	10		
Total reward	27		

(that is, 5 papers with "Yes" bid, 1 paper with "Maybe" bid, and 2 papers with "No" bid).

Start quiz

Take a Quiz!

1. How many papers should you bid on?

- ☐ You may bid on any number of papers.
- ☐ Between 4 and 12 papers.
- ☐ 4
- ☐ 15

2. How many papers will you be assigned?

- ☐ You may be assigned any number of papers.
- ☐ 8
- ☐ 4
- ☐ Between 4 and 12 papers.

3. If you have not submitted bids, what is expected to happen?

- ☐ We will try to assign you the most interesting papers for you.
- ☐ You will receive exactly 8 papers according to our needs.
- ☐ You will not be assigned any papers.

4. if you submitted a large number of bids, what is expected to happen?

- ☐ You will get a higher bonus.
- ☐ Your will get a higher reward.
- ☐ You will receive exactly 8 papers according to our needs.
- ☐ You will receive exactly 8 papers from your selection according to our needs.

5. Paper A is worth 30 bidding points and paper B is worth 65 bidding points:

- ☐ You will get a higher bonus if you get paper B.
- ☐ You are more likely to get paper B if you bid on it.
- ☐ You are more likely to get paper A if you bid on it.
- ☐ You will get a higher bonus if you get paper A.

6. Paper A contains (in the title, keywords or abstract) 2 of your personal keywords and paper B contains 3 of your personal keywords:

- ☐ You will get a higher bonus if you get paper A.
- ☐ You will get a higher bonus if you get paper B.
- ☐ You will get the same bonus in both papers.

7. Following the example in the instructions, after the experiment is over, Jane is assigned 8 papers: {#174, #196, #304, #332, #347, #66, #3, #189}.

Suppose now that paper #174 is highly relevant, papers {#196, #304, #347, #332} are relevant, paper #3 is irrelevant and papers {#66, #189} are without bonus.

Then the total reward to Jane is:

- ☐ 22
- ☐ 23
- ☐ 24
- ☐ 25

[Go back to the instructions](#)

SUBMIT

Figure 9: