Mitigating Skewed Bidding for Conference Paper Assignment

Inbal Rozencweig, Reshef Meir, Nicholas Mattei, and Ofra Amir

Abstract

The explosion of conference paper submissions in AI and related fields 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*, i.e., those papers that receive no bids. In an attempt to understand and mitigate this problem, 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, in particular the order in which papers are presented [10, 15]; and information on paper demand [32]. Our results suggest that several simple adaptations, that can be added to any existing platform, may 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 [5]. In some areas of computer science (mainly AI/ML), 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 [32]. For this matching to happen, the committee members must first submit their preferences over papers. These preferences are supposed to reflect both the competence and the interest of the reviewer 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 [10] provide a detailed account of conference bidding and review flow. After the bidding process, one of the many algorithms for matching under preferences [28, 23] can be used to find an assignment satisfying various notions of optimality, fairness, stability, etc. [12, 1, 2, 3].

Crucially, the current design of the bidding process falls far short of eliciting the full preferences and capabilities of reviewers. First, in some widely used platforms (e.g. EasyChair) there are only three levels of preference: 'no' / 'maybe' / 'yes'. Other platforms provide a finer scale for reviewers to express their preferences. However, it is not clear to what extent reviewers use this flexibility, as *extreme responding* or *scale end bias* is a well known phenomena in many social sciences [17]. Additionally, it is not clear 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, an example being the Toronto Paper Matching System (TPMS) [11].

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 [15, 32, 37, 38, 27, 26]. The argument is that 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 [10] 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![1] A poor assignment, in turn, may affect review quality [40, 35, 36, 25]; and increase the overhead on conference chairs, who need to handle these *orphan papers* that receive no bids via manual (re)assignments. Skewed bidding is also likely to put obstacles in the way of achieving alternative goals such as fairness [34, 27], as creating a fair assignment crucially depends on actually knowing the preferences of the reviewers.

Skewed Bidding at AI Conferences At AAMAS, where we have data from PrefLib [30, 31], there are also a high number of orphan papers.[2] The AAMAS 2015 dataset contains 9,817 bids of 201 reviewers over 613 papers; this represents about 40% of the actual 22,360 bids of 281 reviewers over 670 papers. The 2016 data contains 161 out of 393 reviewers with bids over 442 out of 550 papers. Within this, for AAMAS 2015 papers had 6.9 bids on average, yet there are 30 papers that have no bids at all (5%) and 95 papers that have less than 3 bids (15.4%), while for AAMAS 2016 papers had 6.5 bids on average, but there are 8 papers that have no bids at all (1.8%) and 54 papers with less than 3 bids (12.2%).

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 [32].

1.1 Proposed Solutions

There have been two recent suggestions in the literature to alleviate the problem of skewed bidding: Presenting low-demand papers higher on the list [10, 15]; and providing information regarding paper demand [32].

Interestingly, the first suggestion builds on reviewers' cognitive biases, while the latter exploits their (bounded) rational behavior.

In more detail, Fiez et al. [15] 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 [33]. Primacy effects have been empirically shown to occur in conference bidding data on ConfMaster [10]. The underlying idea is that demand can be smoothed by taking advantage of well known cognitive biases rather than providing more information to bidders.

We should note that by default, most platforms order papers by their submission number. Typically this is a serial number assigned on submission, but recently some platforms such as HotCRP started to assign random submission numbers. In addition, users can usually sort papers according to every column (e.g. alphabetically by title, or by quality of matching according to keywords).

The other suggestion, by Meir et al. [32], considers 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 induces an incentive to follow the recommended instructions, and thereby reduces the skew in bids and leads to an improved assignment. Drawing inspiration from the Trading Post Mechanism [39], they suggest tagging papers with their *inverse price* rather than actual demand, and assign a *budget* the bidder is encouraged to use. Interestingly, rational bidders then have an incentive to exhaust their budget, but some bias in favor of high-price (low-demand) papers

[1] Indeed, 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.

[2] Note that AAMAS reviewers were able to opt out of being included in the public dataset, hence some papers and bids are missing from this dataset.

is necessary to obtain more balanced bids. Thus the model predicts bounded rationality would lead to the best results.

In both the work of Meir et al. [32] and Fiez et al. [15], the actual behavior of the individual bidder (i.e. how their bid is affected by order or 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.

1.2 Contribution

The goal of this paper is to explore how different components of the bidding platform affect the probability that a participant will select a particular paper. The main motivation, following [15, 32] is to promote the selection of papers with few bids, thereby reducing the skew and indirectly improving the paper assignment.

Since previous work has suggested to control either the order of papers [10, 15], or the information given to users on the demand [32], 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 effect on most subjects, but in a rather weak 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.

1.3 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 [24]. In an academic context, papers appearing earlier on an email digest are more likely to be downloaded and cited [14]. Cabanac and Preuss [10] were the first to show that ordering effects occur in paper bidding. Later, Fiez et al. [15] suggested a sophisticated sorting algorithm that takes into account both dynamic demand and estimated reviewers' preferences.

Rodriguez et al. [36] 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 or keyword occurrence. The authors find very low correlation between reviewers' areas and their bids, and conjecture that *reviewer fatigue* may be responsible. Our work does not get into whether reviewers' preferences are indeed based on expertise (as

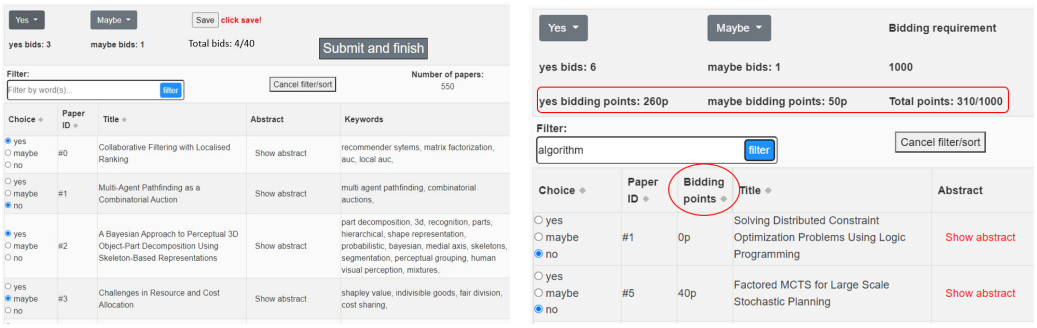


Figure 1: Left: Example of our bidding interface used for all experiments. Right: The interface with the additional column for iPrices (called 'bidding points') and the budget.

opposed to, say, curiosity and interest in the title). It does however shed light on the other, factors that consistently affect bidding behavior.

A major challenge in behavioral studies is having subjects with real-world preferences *and* comparing behavior against true preferences, which are private. Ideally, we would combine these in a single experiment that cleverly elicits the real preferences, as in the work of Budish and Kessler [9] on course allocation, or by performing individual exit polls [4] on voters. Since there is no conference, let alone a large one, that uses a similar mechanism for paper bidding, we resorted to use a combination of field and controlled experiments.

Assignment Algorithms The assignment of papers to reviewers is formally a version of the multi-agent resource allocation problem with capacities [6] and has been well studied in a number of areas of computer science [20, 27], economics [8], and beyond [13]. Garg et al. [18] 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. 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.

Some modern platforms use TPMS or other systems that infer the interests of the reviewer from her list of publications or other sources [11]. However it does not seem that implicit preferences induced from TPMS are less skewed than explicit bids. As Fiez et al. [15] find in their study, TPMS scores result in a very skewed and sub-optimal bid distribution, where many papers receive very low scores. For example, in the TPMS dataset from ICLR 2018, out of the 911 papers, 85 of them (9.3%) have a *maximum* similarity score ≤ 0.1 (on a $[0, 1]$ scale), meaning that these papers are very unlikely to get bids from reviewers.

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 subject is presented with a table containing all papers. For each paper, the table specifies the title and keywords, and 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

[3] See https://easychair.org/ and https://confmaster.net/.

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 their selection of the papers at any time. Subjects could sort papers according to any column and 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. [32] 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 section. Following Shapley and Shubik [39], Meir et al. [32], 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.

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 as in Fig. 1(Left). 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 [32]: (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 [27, 18]. The implementation was the same as that of Lian et al. [27] 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

[4] https://comsoc2021.net.technion.ac.il/

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

be explained below. All treatments are between subjects. All subjects faced the same set of 550 papers from AAAI'15, which are publicly available.[6] Subjects in groups **B**, **H** were requested to bid on 30 papers (40 in some cases), of which 8 will be assigned. Subjects in groups **P** had a budget of 1000 bidding points.

Setting Paper Demand As each subject in the controlled experiment is independent, we needed to generate the demand (i.e. the iPrices) for each paper. Rather than generating artificial demand and derive the iPrice from it, 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 The assignment in the controlled experiment plays no role in our analysis and is described in the full version. Subjects 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). The final assignment was displayed to the subject immediately after they submitted their bid, together with the breakdown of the reward.

Incentives In our controlled experiment, subjects 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 subject 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 as common in MTurk Experiments [29]. An important remark is that in real conferences reviewers' interests are often positively correlated. Using common keywords leads to a similar situation in our controlled experiment with a correlation of 0.7 ± 0.16 in paper relevance among subjects. The personal keywords were selected at random for each subject from the pool of all papers' keywords, with constraints to make sure all subjects had a similar amount of relevant papers. These personal keywords were displayed in a separate box on the screen.

Instructions and Demo To make sure that the (rather complex) instructions of our experiment are understood we: detailed instructions; an online quiz; and a demo game. The study was approved by the IRB of the authors' institution. For further details see the full version.

2.4 Measuring Behavior

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

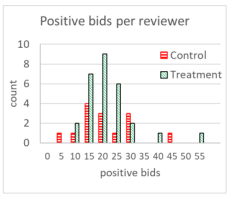
For a set of presented papers S, we denote by $C(S) \subseteq S$ the subset of papers that were selected by subjects. Note that each paper is presented to multiple subjects, and counted as a separate 'presented paper' for each subject. Also note that we treat any positive bid ('Maybe'/'Yes') as a selection. 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.

[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



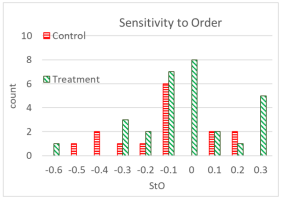


Figure 2: Left: Example of calculating reward and sensitivity parameters. Selected papers are marked with *. The reward is for a subject with the personal keywords {**Red**, **Graphs**, **Algorithms**}. Middle: Histogram of the number of bids for each reviewer in the field experiment. Right: Distribution of individual order sensitivity values in the field experiment.

Measuring Individual Behavior For each measured feature $X \in \{R, O, D\}$ (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(Left) paper #3 has $f^O(s) = \frac{3}{7}$, $f^D(s) = \frac{40}{100}$, and $f^R(s) = \frac{0}{2}$, as the maximal reward in this example is 2.[7]

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 $f^D(S') = \frac{1}{3}(0.3 + 0.8 + 0.4) = 0.5$.

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:

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

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;
- $StO = \frac{6}{14} - \frac{22}{35} = -0.2$, 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).

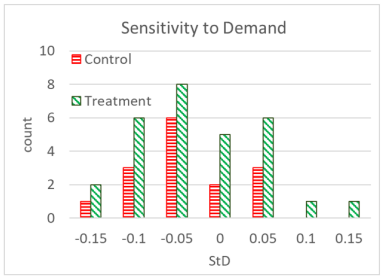
Note that StR cannot be evaluated in the field experiment since we have no direct access to the reviewers' real preferences and expertise.

Measuring Group Behavior One way to measure the group behavior is considering the average StX values of group S members (denoted $StX(S)$). 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 $PS(S') := \frac{|C(S')|}{|S'|}$. We can then test if the behavior in two conditions S, S' is different by comparing $StX(S)$ to $StX(S')$ or $PS(S)$ to $PS(S')$, checking if the different is significant using an unpaired t-test.

3 Results from the Field Experiment

The empirical distribution of bids is shown in Figure 2 (Middle). 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.

[7] The reward scheme we actually used was a bit different. In particular, the reward for papers with 0 personal keywords, which are most papers, was negative, so there is a strong incentive to avoid them. See full version.



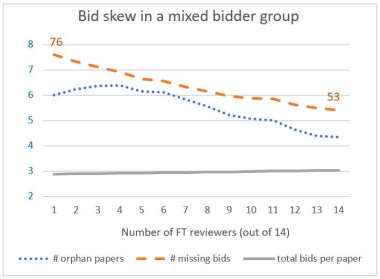


Figure 3: Left: Distribution of demand sensitivity values in the field experiment. Right: Bootstrap results for number of underdemanded and orphan papers in mixed group of 14 reviewers.

To see if the induced bids in both conditions are drawn from different distributions, we used a two sample Kolmogorov-Smirnov test with the null hypothesis that the treatment distribution was less than the control distribution [22]. This resulted in a test statistic of 0.001429 and a p-value of 0.67, so we cannot reject the null hypothesis that average bid amounts are the same.

However what we really want to know is whether bidders where affected by the other factors, in particular order and demand.

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

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. It is more instructive to look at the distribution of StD values (Fig. 3(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.

3.1 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 [7]. 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. 3(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.

3.2 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

Code	Condition	subjects	non-spammers	games	StReward	StOrder	StDemand
B	Base	50	29	29	0.34 ± 0.07	-0.11 ± 0.10	-0.01 ± 0.03#
P	iPrices	124	80	80	0.34 ± 0.04	-0.16 ± 0.07	0.08 ± 0.04
H	Highlight	43	21	39	0.36 ± 0.10	-0.13 ± 0.07	0.05 ± 0.04
PS	P+ Sort	34	17	17	0.29 ± 0.06	–	0.14 ± 0.10
PHS	P+H+Sort	33	28	59	0.44 ± 0.11	–	0.09 ± 0.10#
IR	Imp. Req.	54	36	36	0.36 ± 0.06	-0.12 ± 0.09	-0.01 ± 0.03
Total (controlled exp.)		338	211	260			
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 subjects and played games in each group in the controlled experiment. The right columns show the average sensitivity of each group (non-spammers only) to each parameter, within 2 standard errors. We mark with # results in the controlled experiment that do not statistically differ from 0.

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.

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]).

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.

Implicit Request (IR) This condition was identical to the base condition, except that the bidding requirement did not appear on the screen during bidding.

Data Collection We collected data from 338 subjects on Amazon Mechanical Turk. Subjects were allowed to play up to three times. Subjects were randomly assigned to the base group or to one of the treatment groups. The total number of subjects of each group appears in the second column in Table 1. The threshold for rejection was set at 12 coins (see 'incentives' above). Note that we deliberately collected more data on the Treatment group (in the field experiment) and the **P**rice group, as the other groups cannot be affected by papers' iPrices.

Figure 4: From left: Histogram of individual StO; Conditional selection probability of a paper; same for demand. 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.

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 the full version.

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 0.3-0.4, and is significantly higher than 0.

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. 4, 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. 4 (top right), primacy affects irrelevant and relevant papers alike, where selection probability drops sharply (about 20%) for papers that are not at the top, and then continues to decrease moderately for another 15-20%.

Our findings regarding order effect are largely consistent with those of Cabanac and Preuss [10] from real conferences, and thus support our Hypothesis 1. The added value of our controlled experiments is two-fold: how order effects are distributed across the population; the dependence of order sensitivity (or lack of thereof) on the relevance of the paper.

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. [32] 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 small effect (barely statistically significant). Looking at the distribution of sensitivity to demand in Fig. 4 (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

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

we see is due to a relatively small number of highly sensitive subjects. This corroborates our initial finding from the field experiment, and partly 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 (only 3 of the 21 subjects demonstrated significant bias towards highlighted papers in Fig. 4). In contrast, about third of the subjects who were exposed to iPrices were significantly affected, and the overall bias doubled.

Which Papers are Affected? Ultimately, the goal of the bidding process is to assign papers to relevant bidders. Adding bids (even on underdemanded papers) promotes this goal only if those added bids are indeed on relevant papers. While we saw that this is not achieved by manipulating the order of presentation, we can see that the effect of high iPrices is mainly on papers that are already relevant (Fig. 4, bottom right).

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.

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. We can see in Table 1 that neither group demonstrates significant increase in sensitivity to demand.

We collected more data for these two conditions, showing that in both of them (but mainly in P+H+S) there is a large group of demand-sensitive subjects, and a smaller distinct group with *negative demand sensitivity*. We suspect that this is an artifact of the experiment, where some subjects deliberately pick low-iPrice papers in an attempt to match exactly 1000 points.

5 Discussion

Our combined experiments in bidding behavior show that:

1. Bidding likelihood increases 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. In the full version we also show that presenting the bidding requirement during bidding (rather than just include it in the instructions beforehand) results in much higher compliance.

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 on experimental results 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. Our preliminary analysis shows that subjects exhibit at least some level of consistency but this should be studied more in-depth over longer time periods and with diverse input.

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?

We argue that the answer is yes. While it is clear that the *preferences* of actual reviewers over real paper assignment are very different from those of AMT subjects in our controlled experiment,

it is much more likely that both groups demonstrate the same *behavioral biases and tendencies* in trying to obtain their preferred outcome.[9]

In that respect, our use of AMT is similar to its use in consumer behavior research, where controlled experiments with simulated (rather than actual) purchases are used to complement field studies and deepen understanding [19]. More generally, results from AMT experiments are considered reliable despite some differences in personality traits [21], especially if subjects are filtered based on their comprehension of the task (as we do).

In addition to the above, there is a concern that the number of participants in the field experiment was too small to make conclusive recommendations. Indeed we see this experiment as a first step, or a 'sanity check' of the proposed approach, and wholeheartedly expect more experiments on a larger scale that will validate the results and deepen our understanding.

Critique on paper bidding with iPrices There are several concerns raised by the suggested bidding scheme in [32]. Mostly regarding fair treatment of papers and strategic considerations of bidders (e.g. is it better to bid earlier or later). Meir et al. [32] directly address most of these concerns in the original paper, where their main point is that bidders are free to ignore instructions and behave as they would without demand information, but any bidder that does take this information into account improves the outcome both for herself and for the others.

We can also add that we did not encounter any adverse effects in our field experiment. However we should keep in mind it was in a small scale.

Another possible objection is that automated matching enabled by systems like TPMS makes bidding redundant altogether, or at least less important. That may be true in the future but as shown in [16] (see our Introduction), current automated fit-scores are also highly skewed, and may therefore exacerbate the problem rather than solve it.

Practical Recommendations We believe that adopting the simple market scheme of Meir et al. [32] 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 [10, 15]. 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.

We recommend doing these changes carefully:

- Consult UX/UI experts regarding the best way to highlight papers so as to avoid confusion, choosing the best terms to describe 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 participants and/or in smaller workshops before full adoption.

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

Acknowledgements Reshef Meir is supported by the Israel Science Foundation (ISF; Grant No. 2539/20). Nicholas Mattei was supported by NSF Awards IIS-RI-2007955, IIS-III-2107505, and IIS-RI-2134857, as well as an IBM Faculty Award and a Google Research Scholar Award. Ofra Amir is supported by the Israel Science Foundation (ISF; Grant No. 2185/20).

[9] Note that we restricted our AMT subjects to similar demographics by requiring a university degree.

References

[1] Haris Aziz, Serge Gaspers, Simon Mackenzie, and Toby Walsh. 2015. Fair assignment of indivisible objects under ordinal preferences. *Artificial Intelligence* 227 (2015), 71–92.

[2] 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).

[3] 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.

[4] André Blais, Robert Young, and Miriam Lapp. 2000. The calculus of voting: An empirical test. *European Journal of Political Research* 37, 2 (2000), 181–201.

[5] J Bohannon. 2013. Who's Afraid of Peer Review? *Science* 342, 6154 (2013), 60–65. https://doi.org/10.1126/science.342.6154.60

[6] 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.

[7] Peter Bruce, Andrew Bruce, and Peter Gedeck. 2020. *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media.

[8] 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.

[9] Eric B Budish and Judd B Kessler. 2017. Can Agents' Report Their Types'? An Experiment that Changed the Course Allocation Mechanism at Wharton. *An Experiment that Changed the Course Allocation Mechanism at Wharton (November 12, 2017). Chicago Booth Research Paper* 15-08 (2017).

[10] 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.

[11] Laurent Charlin and Richard Zemel. 2013. The Toronto paper matching system: an automated paper-reviewer assignment system. (2013).

[12] Jiehua Chen, Robert Ganian, and Thekla Hamm. 2020. Stable matchings with diversity constraints: Affirmative action is beyond NP. *arXiv preprint arXiv:2001.10087* (2020).

[13] 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.

[14] 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.

[15] 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.

[16] Tanner Fiez, Nihar B Shah, and Lillian Ratliff. 2020. A SUPER* Algorithm to Optimize Paper Bidding in Peer Review. *arXiv preprint arXiv:2007.07079* (2020).

[17] Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personality and individual differences* 7, 3 (1986), 385–400.

[18] Naveen Garg, Telikepalli Kavitha, Amit Kumar, Kurt Mehlhorn, and Julián Mestre. 2010. Assigning papers to referees. *Algorithmica* 58, 1 (2010), 119–136.

[19] Anindya Ghose, Panagiotis G Ipeirotis, and Beibei Li. 2014. Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science* 60, 7 (2014), 1632–1654.

[20] 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)*.

[21] Joseph K Goodman, Cynthia E Cryder, and Amar Cheema. 2013. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making* 26, 3 (2013), 213–224.

[22] John L Hodges. 1958. The significance probability of the Smirnov two-sample test. *Arkiv för Matematik* 3, 5 (1958), 469–486.

[23] 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/CBO9781107446984.015

[24] 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.

[25] Omer Lev, Nicholas Mattei, Paolo Turrini, and Stanislav Zhydkov. 2023. PeerNomination: A novel peer selection algorithm to handle strategic and noisy assessments. *Artif. Intell.* 316 (2023), 103843. https://doi.org/10.1016/j.artint.2022.103843

[26] Kevin Leyton-Brown, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, Dinesh Raghu, et al. 2022. Matching Papers and Reviewers at Large Conferences. *arXiv preprint arXiv:2202.12273* (2022).

[27] 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.

[28] David Manlove. 2013. *Algorithmics of matching under preferences*. Vol. 2. World Scientific.

[29] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.

[30] Nick Mattei and Toby Walsh. 2013. PrefLib: A Library for Preferences, HTTP://WWW.PREFLIB.ORG. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT'13)*.

[31] Nick Mattei and Toby Walsh. 2017. A PREFLIB.ORG Retrospective: Lessons Learned and New Directions. In *Trends in Computational Social Choice*, U. Endriss (Ed.). AI Access Foundation, Chapter 15, 289–309.

[32] 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.

[33] 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.

[34] Justin Payan and Yair Zick. 2021. I Will Have Order! Optimizing Orders for Fair Reviewer Assignment. *arXiv preprint arXiv:2108.02126* (2021).

[35] 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.

[36] 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.

[37] Nihar B Shah. 2021. Systemic Challenges and Solutions on Bias and Unfairness in Peer Review. *Working Paper* (2021).

[38] 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.

[39] L. Shapley and M. Shubik. 1977. Trade using one commodity as a means of payment. *Journal of Political Economy* 5, 85 (1977), 937–968.

[40] 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.

Inbal Rozencweig,
Technion—Israel Institute of Technology
Haifa, Israel
Email: inbalroz91@gmail.com

Reshef Meir,
Technion—Israel Institute of Technology
Haifa, Israel
Email: reshefm@ie.technion.ac.il

Nicholas Mattei
Tulane University
New Orleans, LA
Email: nsmattei@tulane.edu

Ofra Amir,
Technion—Israel Institute of Technology
Haifa, Israel
Email: oamir@ie.technion.ac.il