Online Appendices

Crowd-judging on Two-sided Platforms: An Analysis of In-group Bias

Alan P. Kwan · S. Alex Yang · Angela Huyue Zhang

Appendix A: Supplemental Tables

A.1. Variable Definition

Table A.1 summarizes the definition of the variables used in the paper.

Table A.1 Variable Definition

Variable Name	Definition
Vote	Binary. 1 represents the vote is in favor of the seller in the dispute; 0 represents the vote is in favor of the buyer in the dispute.
Seller	Binary. 1 represents the vote is cast by a juror who is registered as a seller in the platform, and 0 represents as a buyer.
Gender	Binary. 0 represents the juror is female, 1 represents the juror is male.
Geographic region	Jurors with known geographic region are classified into five categories: east, west, central, northeast, outside mainland (including Hong Kong, Macau, Taiwan, and overseas). The classification follows the instruction according to the National Bureau of Statistics of China published in 2016.
Age	Age of the juror at the time of the vote.
Experience point	Juror's cumulative experience points earned in the crowd-juror platform at the time the vote is cast.
Experience level	Juror's experience level on Taobao's crowd-juror system (level 1 to level 8) at the time the vote is cast. See Table A.4 for details of the mapping between experience points and experience level.
Case outcome	Binary. 1 represents the final case outcome is in favor of the seller; 0 represents the outcome is in favor of the buyer.
Case duration	The time interval (in minutes) between the time when the case is made available to jurors and that when the case is decided.

A.2. Additional Summary Statistics on Sub-periods

Table A.2 provides the same summary statistics as in Table 1, but for the two sub-sample periods (7-vote majority period and 16-votes majority period) separately. We note that 90% of the cases are determined during the 7-votes period (562,092 vs. 56,237). The percentage of votes in favor of the sellers is comparable between these two sub-periods. In both periods, juror experience is highly skewed, and most votes are contributed by experienced jurors.

Table A.2 Summary Statistics on the Two Sub-Periods

Panel A: 7-Vote Majority Period

	N	Mean	Std. Dev.	25th	Median	75th
% of vote in favor of seller (by case)	562,092	40.75	32.65	12.5	36.36	70
Number of cases decided (by juror)	128,408	39.75	551.07	1	3	8
Experience points (by juror)	128,408	2974	23007.8	50	120	432
Experience points (by vote)	5,104,267	239,720	256,002	10,682	153,345	404,225

Panel B: 16-Vote Majority Period

	N	Mean	Std. Dev.	25th	Median	75th
% of vote in favor of seller (by case)	56,237	39.29	32.72	11.11	27.27	69.57
Number of cases decided (by juror)	36,141	31.49	1456.35	2	4	10
Experience points (by juror)	36,141	9,703	40,582.91	114	390	1514
Experience points (by vote)	1,138,048	148,457	186,502	10,140	87,960	215,535

Notes. Row 1 reports the % vote for sellers wherein the unit of observation is a case. Row 2 reports the number of cases decided per user, on average. Row 3 reports experience points at the juror level (averaged over all the votes they cast during the sample period). Row 4 reports experience points at vote level.

Table A.3 provides summary statistics related to feedback and vote consistency (see Appendix E for detailed description and analysis), both for the entire period, and for the two sub-period (7-votes and 16 votes). For feedback, $NetOut_{jt}$ is a juror-day level measure, defined as follows.

$$NetOut_{jt} = \frac{N^{case}_{j,t-1} - N^{vote}_{j,t-1}}{N_{j,t-1}}, \tag{6}$$

where $N_{j,t-1}$ is the total number of cases juror j participated on their last active day before day t (let it be $t-1$), $N^{case}_{j,t-1}$ is the number of cases voted in favor of juror j's out-group by the majority, and $N^{vote}_{j,t-1}$ is the number of cases that juror j voted in favor of their out-group.

For vote consistency, $Consistency_{ij}$ is a juror-case level measure. It equals to 1 if juror j's vote on case i is consistent with the majority of the remaining jurors who also voted on case i, and 0 otherwise. We remove the observations when the remaining jurors' votes result in a tie.

A.3. Summary Statistics based on Taobao Experience Level

As mentioned in Section 3, Taobao classifies the crowd-jurors by their experience levels, which ranges from Level 1 to Level 8. Table A.4 summarizes jurors' experience levels with the number of jurors and the number of votes they cast. As shown, while 57% of the jurors belong to Level 1, they cast only 12% of the votes. In comparison, Level 8 jurors, who only account for 0.5% of the total number of jurors, cast more than 4% of the votes. In other words, on average, Level 8 jurors vote more than 30 times more frequently than Level 1 jurors.

Table A.3 Summary Statistics: Feedback and Consistency

Panel A: Full Sample

	N	Mean	Std. Dev.	25th	Median	75th
NetOut (Juror-day level)	292,025	0.004	0.286	-0.091	0.00	0.091
NetOut (Vote level)	5,488,367	-0.015	0.210	-0.111	0	0.069
Consistency	5,488,367	0.762	0.426	1	1	1

Panel B: 7-Vote Majority Period

	N	Mean	Std. Dev.	25th	Median	75th
NetOut (Juror-day level)	212,468	0.001	0.282	-0.091	0.00	0.91
NetOut (Vote level)	4,490,855	-0.018	0.203	-0.111	0	0.070
Consistency	4,490,855	0.755	0.430	1	1	1

Panel C: 16-Vote Majority Period

	N	Mean	Std. Dev.	25th	Median	75th
NetOut (Juror-day level)	79,557	0.014	0.295	-0.111	0.00	0.087
NetOut (Vote level)	997,512	-0.006	0.236	-0.095	0	0.063
Consistency	997,512	0.797	0.401	1	1	1

Notes. $NetOut$ is a measure between -1 and 1 defined at juror-day level. $NetOut$ (Vote Level) presents the summary statistics of the variable by considering how many votes the focal juror cast on that day. Consistency is a binary variable defined at vote level.

Table A.4 Summary Statistics by Experience Levels

Experience Level	Experience Points (x)	# Votes	% Votes by Buyers
1	$0 \leq x < 600$	537,445	78.9
2	$600 \leq x < 4000$	738,708	80.5
3	$4,000 \leq x < 12,000$	325,488	80.0
4	$12,000 \leq x < 54,600$	729,037	83.0
5	$54,600 \leq x < 171,000$	1,093,694	85.7
6	$171,000 \leq x < 390,000$	1,381,549	89.7
7	$390,000 \leq x < 744,000$	1,132,140	84.6
8	$x \geq 744,000$	304,254	92.6

Appendix B: Do Jurors Skip Cases? An Empirical Investigation

As discussed in the main body of the paper, to mitigate the concern that our estimate of in-group bias coefficient (e.g., β in Eq. (1) is biased by jurors strategically selecting certain cases, we have included case fixed effects in most of the specification. As explained in Section 5, by doing so, the coefficient of interest β can be interpreted as *for a given case,* how much more likely a seller juror will vote in favor of the seller than a buyer juror will.

Although case fixed effects alleviates the concern of selection, our inferences would be cleaner if the cases are indeed (effectively) randomly assigned to different jurors. As the Public Jury has institutionalized a multi-layered randomization procedure in the case broadcasting process (detailed in Section 3), which ensures that the cases assigned to each juror are effectively random. Thus, the only concern is that some jurors may strategically skip certain cases allocated to them. Unfortunately, the data regarding juror skipping cases is not retained in the system, so we cannot directly document the frequency of such case skipping behavior. To further alleviate this concern, we took a two-pronged approach.

First, we interviewed Taobao managers in charge of the Public Jury system, who have confirmed that based on their observations, case skipping was indeed rare for two reasons. First, the Public Jury system has implemented a number of monitoring mechanisms to detect behaviors it deemed suspicious. For example, if a juror has skipped many cases in a row, this will raise a red flag from Taobao. If such behavior persists, Taobao will suspend the jurors account for a period of time, or even indefinitely. Such mechanisms significantly increase the cost for jurors who want to game the system (e.g., to find cases that they have an interest to manipulate their outcome). Second, for those jurors who are genuinely interested in judging, Taobao's point and ranking system provides them with strong incentives to complete the cases allocated to them, especially considering that there is an abundant supply of daily active jurors relative to the number of cases for these jurors to work on. This further discourages jurors to skip cases as they lose the opportunity to advance their ranks. It is thus highly unlikely for this type of judges to skip a case after having invested time to read the case materials.

Second, we conduct an empirical analysis on the possibility of case selection/skipping to the best of our ability given the data limitation. We note that a common approach to test for case selection in similar settings is to conduct a "balancing test", that is, to examine whether juror characteristics are correlated to case characteristics (e.g., Shayo and Zussman 2011). Unfortunately, as we do not have data on case characteristics, we opt to construct a measure of case heterogeneity by the vote balance decided by the first five votes. For example, if four out of the first five jurors voted in favor of the seller, and 1 in favor of the buyer, the vote balance equals $4-1=3$. This measure allows us to put all cases into 6 groups, with vote balance equals to -5, -3, -1, 1, 3, and 5 respectively. We then calculate the characteristics of the 6th and 7th voting jurors for each of this six case categories. By separating the jurors whose votes are used to categorize the cases and those whose characteristics to be examined, we alleviates the concern that the case categorization is affected by juror characteristics. Further, we note that the concern most relevant to our estimation of in-group bias is that jurors skip/select cases along the line of their buyer/seller status. Specifically, for our estimate of in-group bias coefficient (β) to be explained by case selection cases, the most likely scenario would be that

seller (buyer) jurors selecting cases that they believe the seller (buyer) side in dispute is more likely to win. Consequently, we would expect that there is a larger portion of seller jurors among the cases categorized as an easy seller win (large and positive vote balance by the first five votes) than among those cases identified as an easy buyer win. Put differently, we would expect a positive correlation between vote balance and seller juror proportions.

Table B.1 Juror characteristics by Vote Balance of the First N Votes During the 7-votes Period

Panel A: $N = 5$ (Vote Balance Calculated by 1-5th Votes, Juror Characteristics of 6-7th Votes)

Vote Balance (Seller - Buyer)	Number of Cases	Seller (%)	Female (%)	Average Age	Avg. Experience Points	Avg. Experience Points (Demeaned)
-5	135,302	16.29	34.50	37	220,773	365
-3	121,326	16.39	35.05	37	215,193	-2,426
-1	92,597	16.26	35.68	36	213,388	-4,787
1	74,720	16.30	35.86	36	216,626	-3,788
3	70,211	15.83	36.25	37	233,723	1,793
5	67,936	15.00	35.83	37	273,614	12,444

Panel B: $N = 3$ (Vote Balance Calculated by 1-3rd Votes, Juror Characteristics of 4-7th Jurors)

Vote Balance (Seller - Buyer)	Number of Cases	Seller (%)	Female (%)	Average Age	Avg. Experience Points	Avg. Experience Points (Demeaned)
-3	196,634	16.06	34.69	37	225,175	-1,185
-1	148,181	16.03	35.47	37	220,347	-3,928
1	111,226	15.77	35.86	36	228,971	-1,127
3	106,045	15.00	36.01	37	266,409	8,870

We do not observe such a relationship in our empirical results summarized in Table B.1. Panel A presents the result where the cases are classified by the vote balance of the first five votes, and the juror characteristics are the average of the 6th and 7th voters. As shown, among different case categories, the proportions of seller jurors are similar. In fact, if any, it appears that among cases judged as a strong seller win (Vote Balance = 3 or 5), the sixth and seventh votes are marginally less likely to be cast by *seller* jurors. For such a pattern to be consistent with juror selecting cases, one possibility is that seller jurors may choose to skip cases that they deem as easy seller wins in order to leave their time and energy to cases that they believe need their vote for the seller side to win. If true, then such behavior actually suggests that the coefficient β is under-estimated, suggesting that in-group bias is probably stronger than our main estimates in Table 2. For robustness, we also classified the cases based on the first three votes, and examined the juror characteristics based on the average of the 4th – 7th voters. The results are presented in Panel B. As shown, we do not observe any pattern along the line of buyer/seller status either.

In summary, while we cannot empirically exclude the possibility that case skipping exist, the evidence we have provides further assurance that our empirical findings in the paper are most likely a reflection of the existence of in-group bias, instead of jurors strategically choosing cases.

Appendix C: Empirical Results on Juror Participation

In this Appendix, we provide the empirical findings on several dimensions of juror participation behavior, including response speed, attrition, and jury composition in response to panel size and case load. These findings not only help us better understand crowd jurors' behavior beyond voting, but also allow us to build more realistic features into our simulation model, and thus generate useful managerial insights. We summarize the main findings as follows.

Response Speed. Across jurors, more experienced ones are faster at responding to cases. As a juror's experience increases, their response speed also increases.

Participation and Attrition. Jurors with more experience, more recent participation, and with more votes aligned with the final case outcomes are more likely to continue participating in the Public Jury. Jurors with more experience exhibit a significantly lower attrition rate.

Juror Behavior in Response to Panel Size and Case Load Shocks. On average, a larger panel size or a higher case load results in a decrease in average juror experience.

C.1. Juror Response Speed

We first examine how fast jurors respond to a case broadcast to them. Our dataset includes two time related items: the time when the case is submitted to the Public Jury for judging, and the time when a juror casts their vote on this case. As described in Section 3, the Public Jury dispatches cases in a batch process. That is, a case may not be broadcast to jurors immediately after submission. Further, due to the randomization procedures that Taobao implements, a case may appear in different positions in a task pack for different jurors. Considering these complications, we organize the time related data as follows. First, we define the item "Waiting Time until First Vote" as the time elapsed between the case submission time and the time when the first vote is cast. This term captures how long a case needs to wait in the Public Jury system before receiving the first vote. Second, we define "Vote Time Span" as the time difference between the first vote and the last vote for a specific case. Finally, the term "Case Resolution Time" is the sum of the above two, capturing the entire time span from case submission to resolution. The summary statistics of these metrics are presented in Table C.1. Over the entire sample period, a case waits for more than 5 hours (336 minutes) on average in the system before receiving the first vote, and takes less than 3 hours (169 minutes) to collect votes. In total, the average case resolution time is less than 9 hours (504 minutes), and more than 75% of cases are resolved within 18 hours. By examining the 7-vote and 16-vote sub-periods separately, we observe that it costs a case in the 16-vote period 6 hours on average to collect votes to meet the required majority, which is 140% more than that for a case during the 7-vote period.

Next, we examine how more experienced jurors behave in terms of response time. One observation we have made in Figure 2 in the main body of the paper is that for a specific case, the average experience point of jurors who cast the earlier votes is higher than those who cast the later votes. This alludes to the fact that more experienced jurors respond faster to a case. We confirm this pattern with a more formal econometric test with the following specification:

$$\log(\text{Response Time}_{ijt}+1) = \alpha + \beta \times \log(\text{Experience Points}+1) + \delta_i + \theta_j + \epsilon_{ijt}. \tag{7}$$

Table C.1 Summary Statistics: Response Time (in minutes)

Panel A: Entire Sample ($N = 618,329$)

	Mean	Std Dev	25th	Median	75th
Waiting Time until First Vote	335.78	469.23	0.60	6.53	676.22
Vote Time Span	168.70	335.95	6.48	45.55	200.10
Case Resolution Time	504.48	627.15	7.12	73.17	1124.70

Panel B: 7-Votes Period ($N = 562,092$)

	Mean	Std Dev	25th	Median	75th
Waiting Time until First Vote	307.63	453.98	0.58	3.32	491.32
Vote Time Span	149.53	236.22	6.08	30.65	183.90
Case Resolution Time	457.18	565.95	6.70	40.25	1015.78

Panel B: 16-Votes Period ($N = 56,237$)

	Mean	Std Dev	25th	Median	75th
Waiting Time until First Vote	617.05	524.97	1.52	791.20	1138.98
Vote Time Span	360.12	801.82	47.00	175.23	469.20
Case Resolution Time	977.18	936.65	49.80	1329.93	1380.05

We define Response Time$_{ijt}$ as the time (in seconds) elapsed from the time when the first vote on case i is cast to the time when juror j votes on this case. By this definition, if juror j casts the first vote, then Response Time $= 0$. By defining response time this way, we ignores the "Waiting Time until First Vote" component, which is mostly due to Taobao's internal policy instead of juror heterogeneity. We also consider a combination of the case effect effects (δ_i) and juror fixed effects (θ_j).

Table C.2 Juror Response time and Experience

	log(Response Time + 1)		
	(1)	(2)	(3)
Intercept	9.28***		
	(0.101)		
log(Experience Points + 1)	−0.130***	−0.0324***	−0.0242***
	(0.012)	(0.0011)	(0.0039)
Case FE		✓	✓
User FE			✓
Observations	6,240,035	6,240,035	6,240,035
R-squared	0.0121	0.952	0.955

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case are reported. Response Time defined as the time elapsed from when the first vote is cast until the current focal vote is cast.

Table C.2 reports our results. Columns 1, 2 and 3 vary the fixed effects, with Column 1 without any fixed effects, Column 2 with only case fixed effects, and Column 3 with both case and juror fixed effects. In all specifications, the coefficient on experience is negative and highly significant. This suggests that in general, more experienced jurors tend to cast their votes faster than less experienced ones. There are two possible reasons. First, experienced jurors may simply process cases faster. Second, experienced jurors may also check in the Public Jury system more often, and thus are able to react to cases faster once broadcast. Unfortunately, as we do not have the data on when a juror starts to process a particular case, we cannot distinguish these two reasons.

Despite the above limitations, the relationship between juror experience and response speed imposes an important implication: "the wisdom of the crowd" in general implies that a larger crowd is associated with higher decision quality. However, in our setting, as our previous results have shown, less experienced jurors exhibit a greater degree of bias. Thus, by increasing the panel size, we face a trade-off: on the one hand, increasing panel size helps better eliminate idiosyncrasies in individual decision-making, thus improving judging quality. On the other hand, a larger panel is associated with, on average, a lower quality juror pool, thus exacerbating bias. As shown in the simulation in Section 6, which builds in the relationship between juror experience and response time (using Column 1 in Table C.2), this severely limits the capability of using a larger panel to mitigate the impact of in-group bias on case outcomes.

C.2. Juror Participation and Attrition

Next, we try to understand juror participation and attrition behavior on the platform. To be clear, we cannot observe whether a juror permanently leaves the Public Jury platform. Instead, we could only observe whether a juror who has participated at least once on the platform during our sample period vote again. Thus, we cannot perfectly distinguish whether a juror's lack of participation is because they have left the system ("attrition") or simply because they respond too slowly relative to other jurors and thus would not be able to cast a vote. Given this limitation, we conduct two separate analyses to jointly examine jurors continued participation and attrition behavior.

In the first analysis, we explore the determinants for juror's continued participation based on a juror-quarter panel. A juror enters a panel the first quarter after they are first observed in sample. They remain in the sample thereafter. We define the dependent variable $Participation_{jt}$ as whether juror i with enrollment date before quarter t participated in any case in quarter t. We choose the longer time unit because over a short period of time, jurors who want to participate may not be available or cases may not be available, but over such a long period of time such as a quarter, a juror who wants to participate should be able to participate in at least one case. We consider the following specification:

$$\begin{aligned}\text{Participation}_{jt} \times 100 = \beta_1 Seller_j + \beta_2 \times \log(\text{Exp Points}_{jt} + 1) + \beta_3 \times \log(\#\text{Case Judged}_{j,t-1}) \\ + \beta_4 \times \log(\#\text{Case Judged "Correctly"}_{j,t-1}) + \eta_t + \epsilon_{jt},\end{aligned} \tag{8}$$

We focus on four variables of interests: the buyer/seller status, juror experience at the beginning of quarter t, the number of cases the juror voted during the last quarter, which captures how active the juror has recently been, and finally, the number of cases judged "correctly" by the juror over the last quarter. A case marked

as "judged correctly" by a juror is defined as the juror makes the same vote as the final outcome, which jurors in the end observe. All covariates of interest are standardized (to mean 0 and standard deviation 1) so their coefficients can be compared directly. The 613,786 observations refer to those jurors who have at that quarter served in at least one case in our sample by the end of the prior quarter. In other words, we do not count participation of jurors not yet observed in our data.

Table C.3 Determinants of Juror Continued Participation

Dependent Variable: Participation × 100				
	(1)	(2)	(3)	(4)
Seller	1.345***	1.781***	1.672***	0.760***
	(0.120)	(0.105)	(0.103)	(0.0949)
log(Exp Point)	6.295***			4.316***
	(0.0745)			(0.0473)
log(# Cases Judged from in the Previous Quarter)		7.841***		1.638***
		(0.079)		(0.0186)
log(# Cases Judged "Correctly" in the Previous Quarter)			7.916***	4.792***
			(0.0764)	(0.200)
Time Fixed Effect	✓	✓	✓	✓
Observations	613,786	613,786	613,786	613,786
R-squared	0.134	0.163	0.168	0.201

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors clustered by user are reported.

We report this analysis in Table C.3. Column 1 relates to experience. More experience in the past predicts more experience in the future, all else being equal. Column 2 says that the continued participation is increasing in the number of cases that the juror voted in the prior quarter. While this might suggest that jurors who simply choose to participate this quarter are likely to want to do so next quarter, it could alternatively mean that jurors who are assigned more cases quasi-randomly might be more inclined to participate. Column 3 looks at jurors who judge cases *"correctly"*. Finally, Column 4 presents all covariates. The number of cases attenuates relative to the other two variables, experience and cases judged correctly, while cases judged correctly having the largest economic magnitude. There are two possible explanations to this result: 1) selection: it could be that jurors who "choose case outcomes more correctly" are more committed to judging, and thus are more likely to continue participating. 2) reinforcement: if a juror enjoys choosing the "correct" outcome, then s/he is encouraged to continued participation when s/he is more aligned with the "correct" outcome. While we cannot distinguish between these two channels, we note that either way, jurors who continue participating on the platform are likely to be those that cast high-quality votes, thus providing some assurance on the overall judging quality on the crowd-judging platform.

In our second analysis, we focus on estimating the juror attrition rate at an aggregate level. As noted above, one challenge in estimating juror attrition is that we do not observe when jurors (effectively) left

the Public Jury. Instead, we only observe whether a juror participated over a given period. For example, if out of 30,000 jurors who have enrolled by Day 1, 750 participated at least once over the next month (Day 1 – Day 30), and 480 participated over the following month (Day 31 – Day 60), we say the participation rates over two months are 2.5% and 1.6% respectively. We note that these participation rates are driven by two factors: case availability (a juror is still in the system, but has no case to vote on) and juror attrition. Thus, if we assume that case availability is stationary over time, the ratio between the participation rates over the two consecutive periods ($64\% = \frac{1.6\%}{2.5\%}$) provides an approximation of the survival rate of these jurors from the first month to the second. We then convert the monthly survival rate to daily attrition rate $1.47\% = 1 - (0.64)^{(1/30)}$.

To implement this approach with juror heterogeneity, at every day over our sample period (Day t), we calculate all of the jurors who were enrolled up until time t. Then, we calculate the number of participants in each experience bucket (Taobao's levels 1-8) the following month and the second month after. Using the previous example, assume that out of the 30,000 enrolled jurors, 25,000 belong to Experience Level 1. Out of these jurors, if only 500 participated in the first month, and 250 in the second month, then the first month participation rate is 2%, and the second month one is 1%. Taking the difference between these two participation rates, we estimate that the survival rate for juror with Experience Level 1 is 50% ($= \frac{1\%}{2\%}$). Similarly, if only 50 jurors out of the 30,000 are at Taobao Experience Level 8, if the participation rates are both 98% over the next two months, the corresponding survival rate is then 100%. We repeat this procedure over our sample period, and then take the average survival rates over time, and then convert the average survival rate to daily attrition rate, which are summarized in Table C.4.

Table C.4 Attrition rates by experience group

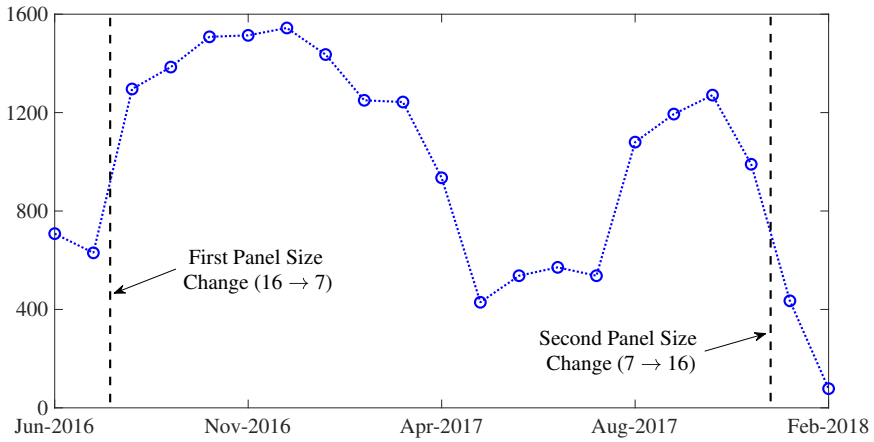
Experience Level	Survival by Month (%)	Daily Attrition Probability (%)
1	51.46	2.190
2	62.43	1.558
3	74.73	0.966
4	86.00	0.502
5	92.45	0.261
6	95.82	0.142
7	98.46	0.052
8	99.05	0.032

Notably, the attrition rate among the more experienced jurors is much lower than that among the less experienced ones. For example, the attrition rate of Level 1 jurors is more than 68 times that of Level 8 jurors. We note that while these attrition rates are estimated based on a number of assumptions, they provide evidence that is consistent with our first analysis regarding continued participation. Further, these estimations are used as the basis to build attrition into our simulation, allowing us to capture more realistic juror participation dynamics. We provide more details in Appendix F.

C.3. Juror Behavior, Panel Size, and Case Load

We now turn to study how jury composition and juror participation behavior vary between the two sub-periods with different voting rules: one period where the platform implements the 7-votes majority voting rule (the 7-votes period, August 2016 – December 2017) and the other with a 16-votes majority voting rule (16-votes period, June 2016 – July 2016 and January – February 2018). As discussed in Section 3, the Public Jury provides crowd-judging as an internal service to different business lines with dispute resolution needs ("internal clients"). These internal clients decide what cases to bring to the Public Jury and also influence the voting policies that apply to the cases they provided. In our sample, these two changes are based on different rationales: the first change, which took place in August 2016, reduced the required majority votes from 16 to 7. At that time, the internal client believed the Public Jury to be a promising mechanism of dispute resolution, and wanted to bring more cases into the Public Jury system. This can be seen in the changes in the daily number of transactional dispute cases in our sample, as visualized in Figure C.1: the daily number of cases jumped from 630 in July (before the panel size change) to 1,296 per day in August (after the change). To meet the increasing demand for jurors, the Public Jury system decreased the required majority.

Figure C.1 Daily number of cases over time (averaged by month)



Regarding the second change, which took place in January 2018 and increased the required majority votes from 7 back to 16, the internal client had decided to reduce the number of dispute resolution cases submitted to Public Jury due to some concerns of the Public Jury system based on the feedback from vendors (including the concern about in-group bias). This can be directly observed in our sample: the daily inflow of cases dropped from 990 in December 2017 to 435 cases in January 2018. This results in a "case shortage", that is, many jurors log in the system finding no cases to vote on. To provide more voting opportunities for jurors, as well as to improve judging quality, the Public Jury decided to increase the majority vote requirement from 7 back to 16. Subsequently, the internal client decided to completely remove this type of cases from Public Jury in February 2018, which also marks the end of our sample period.

As these two panel size changes are both associated with other changes on the platform (e.g., caseload), they are not clean experiments. Thus, in the main body of the paper, our focus is to ensure that our main inference was robust in each of these sub-periods. This is indeed what we found. We present some of the sub-sample results in the main body of the paper (e.g., Columns 3–6 in Table 2), and others in Appendix D. Further, we attempt to conduct additional analysis to shed light on the implication of panel size changes.

We hypothesize this change in voting rules affects juror behavior in two ways. First, keeping the number of cases constantly, an increase in required vote number per case increases the overall demand for jurors. As shown above in the response speed analysis, this demand surge could result in an exhaustion of more experienced jurors, and thus lowering the average experience of the voting jurors. This force increases in-group bias. Second, jurors may in general behave differently under different panel size requirements. We have tentatively examined this possibility in Table 2 (Columns 7 and 8), where we show that by looking at only the first 7 votes, individual juror in-group bias does not differ significantly between the two sub-sample periods. To further investigate this issue, we conduct two more analyses.

First, we focus on the impact of panel size change on jury composition. As noted above, the two panel size changes are both coincident with changes in case load. In fact, Figure C.1 reveals that there exist substantial variations in case load over the entire sample period. Thus, together with panel size changes, we also consider caseload variation to control for temporal mismatch between case load variation and jurors capacity. The specification we use is as follows.

$$\begin{aligned} LogExpMedian = \alpha + \beta_1 \times LargePanelFirst + \beta_2 \times LargePanelSecond \\ + \beta_3 \times CaseLoadGrowth + TimeTrend + \epsilon. \end{aligned} \tag{9}$$

We run two regressions: a day-level time-series regression and a case-level one. The dependent variable $LogExpMedian$ captures the logarithm of the experience point of the median juror on a day (for the day-level regression, $LogExpMedianByDay$) or in a case (for the case-level regression, $LogExpMedianByCase$). The source of variation is in the time-series: $LargePanelFirst$ is the dummy variable for whether the platform is currently in the first 16-vote majority period (June – July 2016), and $LargePanelSecond$ indicates the second 16-vote majority period (January – Feburary 2018). $CaseLoadGrowth = \log(\frac{cases_t + cases_{t-1}}{\sum_{k=2}^{11} cases_{t-k}})$, capturing the localized change in case load. Further, to take into account factors such as experience growth over time, we also control for a linear time trend, that is, $TimeTrend = 1$ if it is the first day of or sample, and it takes a value of 2 if it is the second day, and so on.

The results are presented in Table C.5. We find strong evidence across all samples and specifications that higher demands for votes (either through increasing panel size or case load) are associated with lower experience for the median juror experienced jurors. We note that this result is also consistent with our prior finding regarding response time: as more experienced jurors tend to respond faster to cases, they cast their vote early. When the number of required votes are low (e.g., small panel size, or low caseload), the inexperienced jurors have limited opportunities to participate in voting. However, as the demand for votes increases, the capacity of these experienced jurors is more likely to be exhausted, and thus leaving more opportunities for more inexperienced jurors.

Table C.5 Panel Size, Caseload, and Juror Experience

Dependent Variable	LogExpMedianByDay			LogExpMedianByCase		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	11.80***	12.73***	14.57***	2.477***	22.75***	31.48***
	(0.091)	(0.273)	(0.293)	(0.230)	(0.231)	(0.262)
LargePanelFirst	−0.459***		−0.720***	−0.161***		−0.798***
	(0.138)		(0.140)	(0.0085)		(0.0093)
LargePanelSecond	−0.869***		−1.625***	−0.468***		−1.408***
	(0.143)		(0.154)	(0.0139)		(0.0142)
CaseloadGrowth		−0.153***	−0.459***		−0.953***	−1.312***
		(0.0431)	(0.0465)		(0.0056)	(0.0062)
Time Trend	✓	✓	✓	✓	✓	✓
Sample	Day	Day	Day	Case	Case	Case
Observations	630	619	619	618,329	611,163	611,163
R-squared	0.089	0.021	0.207	0.0056	0.0474	0.0724

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively.

Our second test directly examines the relationship between in-group bias and panel size. We adopt the similar specification in our base regression (Eq. 1), but adding panel size changes as an indicator variable. The specification we consider is as follows:

$$\begin{aligned} VoteSeller_{ijt} \times 100 = {} & \beta_1 \times Seller_j + \beta_2 \times LargePanelFirst + \beta_3 \times LargePanelSecond \\ & + \beta_4 \times Seller_j \times LargePanelFirst + \beta_5 \times Seller_j \times LargePanelSecond \\ & + \gamma_1 \times LogExp_{jt} + \gamma_2 \times Seller_j \times LogExp_{jt} + \delta_i + \theta_j + \epsilon_{ijt}, \end{aligned} \tag{10}$$

where $LargePanelFirst$ and $LargePanelSecond$ are the same dummy variables as defined above. The coefficients of interest are β_4 and β_5. They reflect whether the magnitude of in-group bias during the 16-vote period is different from that during the 7-vote one. We also control for juror experience ($LogExp_{jt}$ represents the logarithm of the juror j's experience point at the beginning of day t plus 1).

The results are presented in Table C.6. In general, we find that the relationship between in-group bias and panel size is unstable across different combinations of fixed effects. For example, after controlling for experience, the coefficient of $Seller \times LargePanelFirst$ is positive and significant with only case fixed effects (Column 4), negative yet insignificant with only juror fixed effects (Column 5), and negative and significant with both juror and case fixed effects (Column 6). The sign of the coefficient around the second panel size change is consistently negative, yet it is not significant under some specifications. If we focus on Column 6, which arguably is our preferred specification, we find that increasing panel size is associated with a decrease in in-group bias. One possible interpretation is that when the panel size is large, a juror under a larger panel size finds it more difficult to influence the case outcome using their vote, and thus vote in a less biased way. However, as the current empirical evidence is unstable under different specifications, we believe more in-depth analysis based on a cleaner empirical setup is required to make a more conclusive claim on this relationship.

Table C.6 In-group Bias and Panel Size Changes

	Dependent Variable: VoteSeller ×100					
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	3.86***			23.34***		
	(1.24)			(2.89)		
LargePanelFirst		−8.98***			−6.08***	
		(0.423)			(0.393)	
LargePanelSecond		23.69***			18.95***	
		(1.11)			(1.05)	
Seller × LargePanelFirst	3.51***	0.064	-0.634	2.04**	−1.40	−1.64**
	(1.12)	(0.786)	(0.584)	(0.968)	(0.919)	(0.754)
Seller × LargePanelSecond	−5.89**	−3.34	−8.17***	−5.20**	−0.812	−6.72***
	(2.53)	(3.74)	(2.99)	(2.55)	(3.29)	(2.53)
logExp				−0.161	6.64***	−0.510*
				(0.145)	(0.500)	(0.277)
Seller × logExp				−1.89***	−3.65**	−1.98*
				(0.379)	(1.57)	(1.08)
Case FE	✓		✓	✓		✓
User FE		✓	✓		✓	✓
Observations	6,242,315	6,242,315	6,242,315	6,242,315	6,242,315	6,242,315
R-squared	0.380	0.161	0.476	0.381	0.164	0.476

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

Appendix D: Robustness Checks

In this section, we conduct a number of robustness checks to solidify our inferences. We summarize these tests in Table D.1.

Table D.1 Summary of Robustness Checks

Table Number	Description
Existence of In-group bias	
Table D.2	Logistics Regression
Table D.3	Different clustering methods
Table D.4	Juror demographic characteristics as moderators
Table D.5	Sub-sample analysis during the 16-votes period
In-group bias and case ambiguity	
Table D.6	Alternative sub-samples and measures for ambiguity
Table D.7	Controlling for juror experience
Table D.8	Sub-sample where initial votes are cast by experienced jurors
In-group bias and perceived threat	
Table D.9	Alternative measures for perceived threat and additional controls
In-group bias and experience	
Table D.10	Alternative sub-samples
Table D.11	Separate buyer/seller analysis
Table D.12	Alternative measure for experience: Experience scaled by the initial level
Table D.13	Alternative measure for experience: Taobao Public Jury Experience Level (1-8)
Table D.14	Alternative measure for experience: Number of cases judged in the sample.

D.1. Existence of In-group Bias

D.1.1. Experience of In-group Bias Based on Logistics Regression In the main body of the paper, we rely on OLS to avoid the incidental variables problem associated with the large number of fixed effects we include in the model. In Table D.2, we report our baseline results on the existence of in-group bias based on logistics regression. As shown, the estimation of in-group bias remain robust for across cases (Column 1), within cases (Column 2), and different sub-samples (Columns 3 – 6).

D.1.2. Existence of In-group Bias under Different Clustering Methods In the main body of the paper, we report standard error double clustered by user and case level. In Table D.3, we show that relative to clustering only at case level (Column 1), this clustering method (Column 2) generates conservative standard errors. Further clustering at the day level (Column 3) does not significantly affect the estimation of standard error.

D.1.3. In-group Bias and Juror Demographic Characteristics Our summary statistics show that jurors are diverse in terms of age and geographic locations (at province level), Geographically, 47.4% of votes are cast by jurors from five coastal provinces, including Guangdong, Zhejiang (where Alibaba is located), Shandong, Jiangsu, and Shanghai.

Table D.2 Existence of In-group Bias Based on Logistics Regression

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	0.202***	0.260***	0.244***	0.350***	0.260***	0.327***
	(0.0547)	(0.0704)	(0.0797)	(0.0533)	(0.0727)	(0.0498)
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE		✓	✓	✓	✓	✓
Sample Period	Full	Full	7-votes	16-votes	7-votes	16-votes
Votes	All	All	All	All	First 7	First 7
Observations	6,242,315	6,242,315	5,088,674	1,153,641	4,686,840	408,954
Psuodo R-squared	0.0488	0.199	0.180	0.332	0.236	0.247

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

Table D.3 Existence of In-group Bias under Different Clustering Methods

	Dependent Variable: VoteSeller ×100		
	(1)	(2)	(3)
Seller	4.36***	4.36***	4.36***
	(0.510)	(1.11)	(1.11)
Date FE	✓	✓	✓
Case FE	✓	✓	✓
Cluster Case	✓	✓	✓
Cluster User		✓	✓
Cluster Day			✓
Observations	6,242,315	6,242,315	6,242,315
R-squared	0.3797	0.3797	0.3797

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively.

We explore whether the existence of in-group bias is mainly driven by a sub-set of jurors. To that end, we augment our baseline specification in Eq. (1) by including different characteristics and their interactions with *Seller*, specifically,

$$\begin{aligned} VoteSeller_{ijt} \times 100 = \beta \times Seller_j + \phi_1 \times Characteristics_j + \phi_2 \times Seller \times Characteristics_j \\ + X'_{jt}\gamma + \eta_t + \delta_j + \epsilon_{ijt}, \end{aligned} \tag{11}$$

where we consider three variables for *Characteristics*: age, gender, and geographic regions.

The results are presented in Table D.4. Our first observation is that across all specifications (considering the third characteristics independently, jointly, and within sub-period), our baseline estimation of in-group bias (β) remains statistically significant. This suggests that the existence of in-group bias is not likely to be solely driven by one sub-group of the juror population, which is consistent with prior with prior research

Table D.4 In-group Bias and Juror Demographic Characteristics

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	4.92***	2.53**	3.58***	2.84**	2.40*	4.58***
	(1.22)	(1.15)	(1.09)	(1.19)	(1.38)	(1.08)
Seller × (Demeaned Age)	0.145			0.227*	0.263*	0.0251
	(0.124)			(0.128)	(0.144)	(0.0952)
Seller × (Demeaned Age)^2	−0.0116**			−0.0148***	−0.0160***	−0.0074**
	(0.0049)			(0.0052)	(0.0058)	(0.0037)
Seller × Female		4.31*		4.93**	5.26*	4.08**
		(2.50)		(2.52)	(2.88)	(1.66)
Seller × (Low Legal Score)			6.15*	4.97	5.21	4.89
			(3.24)	(4.00)	(4.25)	(4.63)
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
Sample	Full	Full	Full	Full	7-votes	16-votes
Observations	6,169,651	6,169,651	5,774,120	5,774,120	4,707,346	1,066,774
R-squared	0.382	0.381	0.385	0.386	0.382	0.401

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors clustered by user and case are reported.

(Hewstone et al. 2002). In addition, we make two observations regarding the impact of juror characteristics. First, we note that in-group bias and juror age correlate in a non-linear fashion, that is, younger and older jurors tend to be less biased than middle aged ones.

Second, we incorporate juror's geographic regions (at province level) into our analysis by mapping each geographic region into the regional legal development score developed by Fan et al. (2018). In Fan et al. (2018), the marketization index on intermediate and legal environment considers a number of factors including the share of public accountants and lawyers in the local population, the quality of the legal environment for businesses as perceived by corporate executives, protection of intellectual property rights in terms of patent applications and research and development grants, and protection of consumer rights. The legal score is granted at the provincial level, with the average score of 6.5 across 31 provinces, and a standard deviation of 4.2 (25th-, 50th-, 75th-percentiles are 3.7, 5,6 and 8 respectively). A higher score indicates that a province has a better legal environment. We create a dummy variable *Low Legal Score* to indicate whether the juror comes from a provincial region whose legal score is among the bottom quantile. We find that those who come from regions with less developed legal systems behave more biased (Column 3). When including other interaction terms, the estimate becomes statistically insignificant, although the magnitude remains similar. This result hints that there could be a connection between the development of the formal legal system and juror bias, which we leave for future study.

D.1.4. In-group Bias: Sub-sample Analysis. Table D.5 presents the results on the existence of in-group bias by further splitting the 16-votes period into two: June–July 2016 (referred to as the 2016 16-votes

period), and January – February 2018 (referred to as the 2018 16-votes period). As shown in Columns (1)–(2), the 2016 sub-period exhibits a statistically significant in-group bias, behaving the same as the full sample and the 7-votes sub-period. However, according to Columns (3)–(4), in-group bias is not significant at the 2018 sub-period. Further investigation reveals that during this period, which accounts for approximately 5% of the total votes in our entire sample, the average seller jurors casting vote are much more experienced than the buyer jurors (average seller experience: 253,064; average buyer experience: 203,752). Further, we note that the total number of votes cast by seller jurors is 24,820. This translate to a mere 8.3% of the total number of votes during that period, while seller jurors contributed 15.4% of the votes during the remaining of our sample period. Finally, we note that this is the period where the Taobao Public Jury gradually winded down this type of cases on the system, eventually removing this case category altogether at the end of February 2018. All these suggest that this period is likely to be an abnormal sub-sample. That said, once we include juror experience (Column 5), we observe that controlling for the difference of judging experience between seller and buyer during this period, the baseline estimate for in-group bias is again positive and statistically significant (15.5%) and this bias is lower among more experienced jurors than among inexperienced ones (the coefficient of the interaction term is -1.59%). Both are directionally consistent with the estimates from the other sub-periods. This provides some assurance that although this period possesses some abnormal features, the existence of in-group bias tends to remain robust.

Table D.5 Existence of In-group Bias during the 16-votes Period

Dependent Variable: VoteSeller ×100					
	(1)	(2)	(3)	(4)	(5)
Seller	6.32***	5.63***	−0.494	−1.50	15.51***
	(0.941)	(0.805)	(2.69)	(2.36)	(4.94)
LogExp					2.38***
					(0.206)
Seller × LogExp					−1.59***
					(0.521)
Controls	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓
Case FE		✓		✓	✓
Sample Period	2016	2016	2018	2018	2018
Observations	853,481	853,481	300,160	300,160	300,160
R-squared	0.0141	0.343	0.0539	0.226	0.226

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

D.2. In-group Bias and Case Ambiguity

D.2.1. Alternative Definitions for Case Ambiguity In Table D.6, we apply a different set of measures to define whether a case is ambiguous. Specifically, in Columns (1)-(4), we define $Ambiguity = 1$ if 1 – 4 votes out of the first five are in favor of the seller. Put differently, the only cases considered non-ambiguous are those that the first five votes are all in favor of the seller or the buyer. In Columns (5)-(6), we classify a case as *ambiguous* if, out of the first 13 votes, 5-8 votes are in favor of the seller. The result shows that among all these alternative definitions for case ambiguity, we continue to observe that jurors exhibit higher in-group bias when facing an ambiguous case relative to a clear-cut one.

Table D.6 In-group Bias and Case Ambiguity: Alternative Definition for Case Ambiguity

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	3.48***		4.02***		4.85***	
	(0.534)		(0.624)		(0.565)	
Seller × Ambiguous	1.13***	0.846**	2.43***	1.81**	3.32***	3.47***
	(0.385)	(0.427)	(0.424)	(0.425)	(1.11)	(1.11)
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE		✓		✓		✓
Sample Period	7-votes	7-votes	16-votes	16-votes	16-votes	16-votes
Votes for Ambiguity	1–5	1–5	1–5	1–5	1–13	1–13
Margin	3	3	3	3	3	3
Votes as DV	6–7	6–7	6–16	6–16	14–16	14–16
Observations	1,124,184	1,124,184	618,607	618,607	168,711	168,711
R-squared	0.662	0.749	0.455	0.564	0.593	0.723

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

D.2.2. Controlling for Juror Experience. Another concern might be that cases predicted as ambiguous may correlate with juror experience. For example, as discussed in Appendix C.3, when the caseload is high, on average, votes are more likely to be cast by inexperienced jurors, who are likely to vote less unanimously on a case (than experienced jurors). Thus, our measure of ambiguity could be correlated with juror experience. To address this issue, in Table D.7, we include $Seller \times LogExp$ as a control. As shown, we continue to observe that $Seller \times Ambiguity$ to be statistically significant for different sample periods and cutoffs for the ambiguity.

D.2.3. Sub-samples where Initial Votes are Cast by Experienced Jurors. One potential concern is that our measure of case ambiguity is correlated with experience among the jurors who cast the initial votes that we use to construct *Ambiguity*. More specifically, if the initial N votes for a case happens to be

Table D.7 In-group Bias and Case Ambiguity: Controlling for Experience

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller × Ambiguous	0.755*	0.994**	0.830*	1.76***	1.76***	3.51***
	(0.391)	(0.453)	(0.427)	(.0420)	(0.497)	(1.11)
Seller × LogExp	−1.17	−1.03*	−1.03*	−2.83**	−2.86**	−2.31
	(0.739)	(0.580)	(0.579)	(1.66)	(1.65)	(1.48)
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE	✓	✓	✓	✓	✓	✓
Sample Period	7-votes	7-votes	7-votes	16-votes	16-votes	16-votes
Votes for Ambiguity	1-5	1-5	1-5	1-5	1-5	1-13
Margin	1	1	3	1	3	3
Votes as DV	6-7	6-7	6-7	6-16	6-16	14-16
Observations	2,248,368	1,124,184	1,124,184	618,607	618,607	168,711
R-squared	0.497	0.589	0.662	0.749	0.455	0.564

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

cast by very experienced jurors, who are better at voting "correctly", then they are more likely to reach a consensus, deeming the case "unambiguous". In contrast, if the initial votes for the case happen to be cast by inexperienced jurors, who tend to vote more randomly, then the case is more likely to be classified as ambiguous. To alleviate this concern, we focus on a sub-sample of cases where the first N jurors are all experienced.

The results are presented in Table D.8. In Columns (1)–(4), we require that the minimum experience points to be 25,000. Columns (1)–(2) study the 7 vote period using the first three votes to classify ambiguous cases. We find similar results as in our main analysis. Columns (3)–(4) study the sixteen vote period, using the first three votes to classify ambiguous cases and the 4th–16th votes to estimate in-group bias. Finally, Columns (5)–(6) define experience as the initial jurors participating having 50,000 points, on average. We find that the estimates of ambiguity on in-group bias to be similar.

D.3. In-group Bias and Perceived Threat

We conduct several robustness checks for our base results on the relationship between in-group bias and perceived threat. First, we note that our measure of threat, $NetOut_{jt}$, includes two types of cases: 1) the cases that juror j voted in favor of his in-group, but the majority rules in favor of the out-group. We refer to this component as $NetOutPos_{jt}$. 2) the cases that the majority rule in favor of juror j's in-group, but the juror voted in the opposite way ($NetOutNeg_{jt}$). By these definitions, it is clear that we have

Table D.8 In-group Bias and Case Ambiguity: Sub-samples where Experienced Jurors Cast Initial Votes

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	3.36***		4.56***		4.28***	
	(0.852)		(0.656)		(0.657)	
Seller × Ambiguous	1.12***	0.757*	1.82***	1.29***	1.73***	1.15***
	(0.360)	(0.392)	(0.406)	(.411)	(0.432)	(0.411)
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE		✓		✓		✓
Sample Period	7-votes	7-votes	16-votes	16-votes	16-votes	16-votes
Votes for Ambiguity	1-3	1-3	1-3	1-3	1-3	1-3
Margin	1	1	1	1	1	1
Min Exp Point for Votes for Ambiguity	25,000	25,000	25,000	25,000	50,000	50,000
Votes as DV	4-7	4-7	4-16	4-16	4-16	4-16
Observations	2,243,393	2,243,393	736,056	736,056	565,291	565,291
R-squared	0.497	0.589	0.449	0.555	0.459	0.568

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

$NetOut_{jt} = NetOutPos_{jt} - NetOutNeg_{jt}$. We then modify the specification in Eq. 4 by using $NetOutPos_{jt}$ and $NetOutNeg_{jt}$ to replace $NetOut_{jt}$. The new specification is as follows.

$$\begin{aligned} VoteSeller_{ijt} \times 100 = {} & \alpha \times Seller_j + \beta_1 \times NetOutPos_{jt} + \phi_1 \times Seller_j \times NetOutPos_{jt} \\ & + \beta_2 \times NetOutNeg_{jt} + \phi_2 \times Seller_j \times NetOutNeg_{jt} \\ & + X'_{jt}\gamma + \eta_t + \delta_j + \theta_i + \epsilon_{ijt}. \end{aligned} \tag{12}$$

The coefficients of interest are ϕ_1 and ϕ_2. As a larger $NetOutPos$ is associated with a higher perceived threat level, and a larger $NetOutNeg$ is associated to a lower level, we would expect ϕ_1 to be positive, and ϕ_2 to be negative.

The results are presented in Columns (1) and (2) in Table D.9. As shown, the sign of our coefficients of interest, namely ϕ_1 and ϕ_2, is consistent with our expectation, and both coefficients are statistically significant, even when we include both case and user fixed effects. Moreover, we note that the magnitudes of these two coefficients are very similar (30.56 vs. 33.05 in Column 2).

Our second extension is related to the observation that our measure for threat, whether it is $NetOut$ or $NetOutPos$ and $NetOutNeg$, includes the focal juror's votes in their last active day. Thus, it is possible that the relationship identified between threat and in-group bias could simply be due to the potential auto-correlation among jurors' vote. To address this concern, we augment the above specification by including the variable $OutVoteLag_{jt}$, which is defined as the number of cases that juror j voted in favor of their out-group

Table D.9 In-group Bias and Perceived Threat: Robustness Checks

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	2.55***		−2.66***			
	(0.772)		(1.11)			
NetOutPos	−41.15***	−18.86***	−30.69***	−13.83***	−0.331	−9.89***
	(0.99)	(1.08)	(1.10)	(1.03)	(0.908)	(0.805)
NetOutNeg	58.14***	14.84***	43.60***	8.62***	−9.91***	3.98***
	(2.49)	(1.03)	(2.17)	(1.06)	(2.25)	(0.781)
Seller × NetOutPos	95.29***	30.56***	96.72***	29.33***		
	(3.44)	(3.00)	(3.87)	(2.94)		
Seller × NetOutNeg	−104.5***	−33.05***	−95.73***	−29.44***		
	(5.59)	(5.32)	(5.54)	(5.52)		
OutVoteLag			16.68***	9.71***	−12.00***	16.33***
			(1.27)	(0.824)	(1.48)	(0.564)
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE		✓		✓	✓	✓
Sample Period	Full	Full	Full	Full	Full	Full
Juror Sample	Both	Both	Both	Both	Seller	Buyer
Observations	5,488,367	5,488,367	5,488,367	5,488,367	797,926	4,690,441
R-squared	0.450	0.490	0.453	0.490	0.738	0.503

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

divided by the total number of cases he participated on his last voting day before time t. Using the notation from Eq. 3 in Section 5.3, we have

$$OutVoteLag_{jt} = \frac{N^{vote}_{j,t-1}}{N_{j,t-1}}. \tag{13}$$

Columns (3) and (4) in Table D.9 present the corresponding result. As shown, even after controlling $OutVoteLag$, both $NetOutPos$ and $NetOutNeg$ remain statistically significant. Finally, Columns (5)-(6) present the separate buyer/seller regressions, again confirming that threat levels appear to affect in-group bias for both buyer jurors and seller jurors.

D.4. In-group Bias and Juror Experience

In this section, we verify the robustness of our results on the relationship between in-group bias and experience by examining various sub-samples and alternative measures for juror experience.

D.4.1. Alternative Sub-samples We present the relationship between in-group bias and juror experience on different sub-samples in Table D.10. As shown in Columns (1)-(2), the relationship between in-group bias and experience points is also statistically significant during the 16-majority-votes period.

Further, as our sample begins a few years after the launch of the crowd-judging program on Taobao, a number of jurors have already accumulated some experience at the beginning of our sample period. Thus,

Table D.10 In-group Bias and Juror Experience: Sub-sample Analysis

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	7.41***					
	(0.93)					
LogExp	2.20***	3.53***	−1.49***	2.83***	−2.61***	3.94*
	(0.267)	(0.858)	(0.364)	(1.03)	(0.482)	(2.24)
Seller × LogExp	−3.56**	−3.36*	−2.04*	−5.62***	−5.60***	−7.49***
	(1.44)	(1.76)	(1.22)	(2.18)	(1.47)	(2.89)
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE		✓	✓	✓	✓	✓
Sample Period	16-votes	16-votes	7-votes	16-votes	7-votes	16-votes
User Active Days	≥ 1	≥ 1	≥ 25	≥ 25	≥ 100	≥ 100
Initial Experience Level	All	All	≤ 4	≤ 4	≤ 4	≤ 4
Observations	1,153,641	1,153,641	1,186,155	270,815	625,259	76,284
R-squared	0.398	0.507	0.650	0.617	0.758	0.753

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

one might wonder if our results were driven by experienced or junior jurors. While this concern does not necessarily change our overall conclusion, it would be helpful to know how robust our findings are across the spectrum of experience. To address this concern, we further restrict our sample based on jurors' experience levels at the beginning of our sample period. Specifically, we focus on jurors whose Public Jury Experience Level is 4 or below at the beginning of the sample period. As shown in Columns (3) – (6), focusing on inexperienced jurors actually makes the results stronger. This is consistent with our conjecture that the correlation between experience and in-group bias is driven by learning, which tends to be most pronounced at the initial stage of learning.

D.4.2. Separate Buyer/Seller Analysis. Similar to the analysis in Columns (5)–(6) in Table 4 (Section 5.3), we separately analyze the impact of experience on buyer and sellers. The results are presented in Table D.11.

As shown, both the full sample (Columns 1 and 2), and the 7-votes (Columns 3 and 4) show that as experience increases, seller jurors consistently vote more for buyers, suggesting that in-group bias is likely to decrease on the seller side. On the buyer side, however, the result is less obvious. Focusing on the 16-votes period, we find that the behaviors during the two sub-period (2016 vs. 2018) are very different: while the 2016 period result is directionally consistent with the 7-vote sub-period, the 2018 period exhibits the opposite results: both buyers and sellers vote more in favor of the seller as they gain experience. However, we note that the result needs to be interpreted carefully for two reasons. First, the sample size during the 2018 16-votes period is very small. Second, from Table D.5, we also know this period behaves in an abnormal way in terms of in-group bias, possible because Taobao is winding down this type of cases in the system.

Table D.11 In-group Bias and Juror Experience: Separate Buyer/Seller Regression

Dependent Variable: VoteSeller ×100								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LogExp	−0.521*	−2.59***	−1.42***	−2.68***	−3.42***	−3.46***	6.47***	13.75***
	(0.278)	(0.640)	(0.279)	(0.545)	(0.456)	(0.986)	(0.994)	(2.27)
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓	✓	✓
User FE	✓	✓	✓	✓	✓	✓	✓	✓
Sample Period	Full		7-votes		16-votes (2016)		16-votes (2018)	
Jurors	Buyer	Seller	Buyer	Seller	Buyer	Seller	Buyer	Seller
Observations	5,300,916	941,399	4,308,501	780,173	717,075	136,406	275,340	24,820
R-squared	0.486	0.711	0.485	0.730	0.455	0.610	0.418	0.696

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

D.4.3. Alternative Experience Measure: Scaled Experience Points Table D.12 presents the result when juror experienced is measured by $\log\left(\frac{\text{Exp Point}_{jt}+1}{\text{Exp Point}_{j0}+1}\right)$, that is, the logarithm of juror j's experience on the day when voting case j normalized by his initial experience at the beginning of the sample period. As shown, the results remain significant under the full sample and different sub-samples.

D.4.4. Alternative Experience Measure: Taobao Experience Level Another alternative measure for experience we consider is Juror Experience Levels defined by Taobao. As discussed in Section 3, Taobao classifies jurors into eight levels, Level 1 to Level 8, with Level 1 being the least experienced juror, and Level 8 being the most experienced ones. The distribution of jurors among these eight levels, and the number of cases they have judged in our sample period is summarized in Table A.4. By interacting the *Seller* dummy with $ExpLevel_{jt}$, juror j's Taobao Experience Levels (1 – 8) when judging cases on day t, Table D.13 reveals that as juror's Taobao Experience Level increases, their in-group bias also reduces.

D.4.5. Alternative Experience Measure: Number of Cases Voted One concern with using jurors' Taobao Public Jury experience points as a measure for experience is that while we have only obtained data from one type of cases – transactional disputes, these jurors could vote on other types of cases that are distributed on the Public Jury system (such as identifying inappropriate languages in reviews). Jurors are rewarded with experience points on all types of cases that they vote on. To isolate experience specific to judging transactional dispute cases, we use the number of cases jurors participated in our dataset (all belonging to transactional disputes) as a proxy for their experience. Specifically, we define $LogNumCase_{ij}$ as the natural log of one plus the total number of cases juror j has voted on starting from our sample period until right before participating on case i.

The results are presented in Table D.14. As shown, with case fixed effects alone, we continue to observe the existence of in-group bias, and that the bias diminishes as jurors decide on more cases. With the addition of juror fixed effects, we observe mild significant result on the main coefficient of interest when we focus on

Table D.12 In-group Bias and Juror Experience: Scaled Experience ($\log\left(\frac{\text{Exp Points}_{jt}+1}{\text{Exp Points}_{j0}+1}\right)$)

Dependent Variable: VoteSeller ×100					
	(1)	(2)	(3)	(4)	(5)
Seller	6.94***				
	(1.29)				
$\log\left(\frac{\text{Exp Points}_{jt}+1}{\text{Exp Points}_{j0}+1}\right)$	0.407	−0.460*	−2.39∗∗	−2.62***	3.94***
	(0.319)	(0.275)	(0.471)	(0.482)	(2.24)
Seller × $\log\left(\frac{\text{Exp Points}_{jt}+1}{\text{Exp Points}_{j0}+1}\right)$	−2.73***	−2.20**	−7.64***	−5.95***	−7.49***
	(0.652)	(1.10)	(2.28)	(1.47)	(2.89)
Controls	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓
User FE		✓	✓	✓	✓
Sample Period	Full	Full	Full	7-votes	16-votes
Initial Experience Level	All	All	≤ 4	≤ 4	≤ 4
Observations	6,242,315	6,242,315	702,083	625,259	76,284
R-squared	0.480	0.380	0.753	0.758	0.752

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

Table D.13 In-group Bias and Juror Experience: Taobao Experience Level (1–8)

Dependent Variable: VoteSeller ×100					
	(1)	(2)	(3)	(4)	(5)
Seller	15.52***				
	(1.82)				
ExpLevel	0.116	0.488*	−1.85***	−2.06***	2.75
	(0.252)	(0.269)	(0.480)	(0.432)	(1.85)
Seller × ExpLevel	−2.54***	−2.05*	−7.87***	−6.11***	−8.48***
	(0.612)	(1.10)	(2.56)	(1.68)	(2.74)
Controls	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓
User FE		✓	✓	✓	✓
Sample Period	Full	Full	Full	7-votes	16-votes
User Active Days	≥ 1	≥ 1	≥ 100	≥ 100	≥ 100
Initial Experience Level	All	All	≤ 4	≤ 4	≤ 4
Observations	6,242,315	6,242,315	702,083	625,259	76,284
R-squared	0.381	0.476	0.753	0.768	0.753

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

Table D.14 In-group Bias and Juror Experience: Number of Cases Judged

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	15.61***		12.61***		10.95***	
	(1.29)		(3.37)		(3.95)	
LogNumCase	−0.440	−0.251**	1.15***	−0.010	−1.27*	−0.524*
	(0.168)	(0.125)	(0.277)	(0.168)	(0.318)	(0.187)
Seller × LogNumCase	−1.88***	−0.404	−1.49**	−0.522*	−1.27*	−0.524*
	(0.377)	(0.266)	(0.510)	(0.305)	(0.668)	(0.316)
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE		✓		✓		✓
Sample	Full	Full	> 100 cases	> 100 cases	> 200 cases	> 200 cases
Observations	6,242,315	6,242,315	5,106,969	5,106,969	4,864,200	4,864,200
R-squared	0.381	0.476	0.432	0.0.497	0.443	0.505

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

jurors who complete at least a certain number of cases (Columns 3–4 for jurors who voted on at least 100 cases over our sample period, and Columns 5–6 for at least 200 cases). Overall, the results are consistent with the main analysis.

Appendix E: Empirical Evidence on Vote Consistency

In the main body of the paper, we focus on in-group bias as the measure for judging quality, and we have shown juror experience helps improve judging quality in this dimension. Of course, ideally, we would quantify judging quality using *voting accuracy*, that is, whether a juror's vote aligns with the objectively correct case outcome. However, as we do not know what the case outcome "should" be, we resort to another proxy, *vote consistency*, that is, whether a juror's vote on a case is consistent with the majority ruling (by other crowd-jurors participating on this case).[26] Intuitively, if the crowd-judging mechanism produces the correct answer, on average, vote consistency could be a reasonable measure for judging quality. Admittedly, another possibility is that a high vote consistency may simply mean that the juror is better at "guessing" what the majority ruling is. However, as there exists no formal interaction between jurors (e.g., panel discussion) in our setting, it is difficult to identify the channel through which jurors directly learn from each other. Moreover, jurors receive no reward nor recognition for correct decisions. In fact, our result in Table 4 suggests that even when feedback is provided, jurors may not necessarily learn to correct their biases. Thus, we think it is more likely that a high vote consistency is a reflection of higher judging quality. Based on this logic and our earlier result on the relationship between in-group bias and juror experience, we predict that as jurors gain more experience, their voting consistency will also increase. Formally, we consider the following specification.

$$Consistency_{ijt} \times 100 = \beta \times LogExp_{ijt} + X'_{jt}\gamma + \eta_t + \delta_i + \theta_j + \epsilon_{ijt}, \tag{14}$$

where $Consistency_{ijt}$ equals 1 if juror j voted consistently with the majority based on the other voting jurors, and 0 otherwise, and $LogExp_{ijt}$ is defined the same as in Table 5. The coefficient of interest is β, which is expected to be positive.

Results in Columns (1)–(2) of Table E.1 confirms our hypothesis. As shown, after controlling case and user fixed effects, the coefficient of experience is not only statistically significant, but also economically meaningful: the increase of vote consistency for a juror from zero experience to median level of experience in our sample (experience points = 133,547) is $0.693 \times \log(133,548) = 8.18$ percentage points, which is equivalent to a more than 10% improvement compared to the unconditional average of vote consistency (77%). We further augment Eq. (14) by including $Seller$ and $Seller \times LogExp$, and the results are presented in Columns (3)–(4). As shown, while seller jurors start with a lower consistency level, they improve significantly faster than their buyer peers as they gain more experience. This hints that the improvement in vote consistency is driven, at least partially, by the reduction of in-group bias. To further investigate whether jurors within the same buyer/seller status also vote more similarly as they gain more experience, we modify the dependent variable by looking at whether a buyer (seller) juror votes in alignment with the majority decision of the remaining buyer (seller) jurors. We refer to this measure as $ConsistencySide$. As shown in Column (5), among only buyer jurors, vote consistency also increases in juror experience. This suggests that as jurors gain experience, they not only converge more with their out-group jurors, but also with their in-group peers. Finally, Column (6) shows that convergence between sellers is not statistically significant. One possible reason is the limited sample size: as only a small fraction of jurors are sellers, and we require at least three seller jurors in one case to construct the $ConsistencySide$ measure, the resulting sample size is much smaller compared to the buyer sample, limiting the statistical power of this test.

[26] We remove observations where other jurors' votes result in a tie.

Table E.1 Vote Consistency and Juror Experience

Dependent Variable	Consistency ×100				ConsistencySide ×100	
	(1)	(2)	(3)	(4)	(5)	(6)
LogExp	1.05***	0.693***	0.923***	0.523***	0.588***	0.014
	(0.0668)	(0.160)	(0.0773)	(0.182)	(0.176)	(0.307)
Seller			−7.15***			
			(1.32)			
Seller × LogExp			0.576***	0.959**		
			(0.168)	(0.415)		
Controls	✓	✓	✓	✓	✓	✓
Month	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE		✓		✓	✓	✓
Juror Sample	Both	Both	Both	Both	Buyer	Seller
Observations	6,242,315	6,242,315	6,242,315	6,242,315	5,294,281	209,678
R-squared	0.249	0.299	0.249	0.299	0.381	0.636

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

E.1. Robustness Tests

Table D.15 shows that in general, vote consistency also increases in juror experience for our two sub-samples (7-vote-majority period and 16-vote-majority period), and consistency among buyers increases more strongly during the 16-vote-majority period.

Table E.2 Vote Consistency and Juror Experience: Sub-sample Analysis

	DV: Consistency ×100				DV: ConsistencySide ×100	
	(1)	(2)	(3)	(4)	(5)	(6)
log(Exp Points + 1)	0.931***	0.355	1.66***	2.42***	0.217	2.36***
	(0.0711)	(0.181)	(0.0849)	(0.480)	(0.198)	(0.485)
Controls	✓	✓	✓	✓	✓	✓
Month	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE		✓		✓	✓	✓
Sample Period	7-votes	7-votes	16-votes	16-votes	7-votes	16-votes
Juror Sample	Both	Both	Both	Both	Buyers	Buyers
Observations	5,088,674	5,088,674	1,153,641	1,153,641	4,301,901	992,380
R-squared	0.265	0.311	0.164	0.255	0.401	0.288

Notes. ***, **, * represent that the estimates are statistically significant at 1%, 5%, and 10% level, respectively. Standard errors double clustered by user and case levels are reported.

Appendix F: Supplemental Technical Details on Simulation

In this appendix, we summarize the technical details of the simulation study discussed in Section 6.

F.1. Modeling Juror Dynamics and Parameter Calibration

We construct a simulation model capturing juror participation and voting behavior, as well as their growth dynamics over 500 days. Next, we describe each component of this simulation model.

Juror Initialization. At day 0, we generate the initial juror pool by including all jurors in our sample who enrolled in Public Jury before June 2016 (the beginning of our sample period. The initial juror pool includes 35,308 jurors, of which 78% are buyers. More than 83% of jurors belong to Taobao Experience Level 1 or 2 (with Experience Points less than 4,000), and only less than 1% of jurors have Experience Level 6 or above.

New Juror Spawning. At the beginning of each day, we assumed there are 190 new jurors (with zero experience) joining the juror pool. The number equals the average number of users enrolled as crowd-jurors during our sample period.

Juror Attrition. We use the estimated attrition rate from Table C.4 as the basis for the daily attrition rate, and scale these rates so that the total number of jurors leaving the platform is similar to the number of new jurors joining every day. The adjusted attrition rates (for each Experience Level) is presented in Table F.1 (Column 2).

Table F.1 Parameters used in Simulation

Experience Level	Daily Attrition Rate (%)	Daily Capacity	Active Rate per Task Pack (%)
1	0.4921	5	0.0255
2	0.3500	8	0.0900
3	0.2171	11	0.2870
4	0.1127	14	0.6116
5	0.0587	19	1.340
6	0.0320	26	4.099
7	0.0116	41	20.57
8	0.0072	71	23.27

Notes. Active Rate per Task Pack applies to jurors who have enrolled in the Public Jury for more than one week. For those enrolled within one week, their Active Rate depends on the number of days that they have enrolled in the system. Specifically, the Active Rate per Task Pack (%) for a juror who joined the Public Jury 0 – 6 days ago are: 20.57%, 0.34%, 0.25%, 0.20%, 0.17%, 0.15%, and 0.14% respectively.

Juror Daily Capacity. To capture the intuition that a juror has a natural capacity for the number of cases they can do on each day, and the observation that jurors with different experience participate in different numbers of cases, we impose a daily capacity for each juror according to their Taobao Experience Level (1–8). The capacity is estimated based on the distribution of the cases jurors from each experience level conditional on that they participate on a day. The estimates used are summarized in Table F.1 (Column 3).

Case Generation. On each day, we generate 1000 cases, which is approximately the average daily number of cases during our sample period. For each case, to account for case heterogeneity, we draw the fixed effect of each case from the empirical distribution of the fraction of votes in favor of the seller in each case. To capture the batch case releasing process in practice, in our baseline scenario, we divide the 1000 cases into 20 batches ("task packs"), each consisting of 50 cases. We conduct a number of sensitivity analyses, including:

1. Uniform case FEs: Drawing case fixed effects from a Uniform[0,1] distribution instead of the empirical distribution of case fixed effects;
2. Random Caseload: Instead of using 1000 cases per day, generate a random number of cases per day according to the empirical distribution of daily case numbers in our sample;
3. Reduced batch size: Changing the daily number of task packs from 20 (each with 50 cases) to 40 (each with 25 cases).

The results of these three scenarios are shown in Table 6 (Rows 2 – 4).

Juror Participation. We take a two-step approach. First, for each "task pack", we randomly draw a number of active jurors according to their Active Rate. In general, the Active Rate is estimated based on the number of participation jurors per day for each experience group and the total number of jurors enrolled in the system adjusted by the attrition rate. The result is presented in Column 4 in Table F.1. One exception is for jurors who have enrolled in the Public Jury recently (within a week). This is based on the observation that compared to jurors with similar experience, a recently enrolled juror is much more active. For example, on their enrollment day, a juror, starting at Experience Level 1, has a more than 75% likelihood of voting on a case on their enrollment day. In contrast, for a Experience Level 1 juror who has enrolled a while ago, the same likelihood is less than 1%. Thus, for those jurors who have enrolled recently, we estimate their Active Rate based on the probability of voting on each of the first seven days after their enrollment date, assuming the enrollment date falls within our sample. For example, if their enrollment date is June 1, 2017, we calculate the probability they participate on every day of the first seven days up until June 8, which then allows us to estimate the probability of participation for a new joining juror.

Second, only considering jurors who still have remaining capacity for the current day, we use the regression results in Table C.2 (Column 1) to generate each juror's response time when facing the case packet based on the juror's experience. We then rank the jurors who are active for this task pack sequentially by their response times.

Juror Voting. Once the juror sequencing is determined, we generate binary random variables to represent whether a juror with certain characteristics votes in favor of the seller or buyer by imputing probabilities using the regression results (Column 1 in Table 5). Recall that the regression estimate suggests that the probability of voting for a seller is related to Buyer-Seller status, experience, and the case fixed effect. We calculate an imputed probability that the juror will vote for the seller. For each case, we draw a parameter β_{case}, which is the fraction of votes for the seller in one of the original cases. For each case, we take the simulated juror's experience and buyer-seller status. From this, we impute the expected probability the juror will vote for the seller based on OLS.

One challenge with our modeling exercise is that we use OLS to account for the incidental parameters problem, but OLS can produce predicted probabilities below 0 or above 1 and assumes that predicted probabilities are linear in the covariates, whereas logistic distributions do not assume this. To make the predicted probability more analogous to what we would get from a binary response model, we apply a correction based on linear discriminant models. Specifically, following Allison et al. (2020), we translate the OLS predicted probability to an estimated logistic distribution probability by applying the linear discriminant model correction.

Voting Policy and Case Outcome. By combining each jurors' votes, their response time (which determines their potential voting order), and the voting rule, we create the case outcome. For example, for the baseline 7-vote-majority policy, we take into accounts each vote according to their response times and reach the outcome of the case once we collect seven votes in favor of one side in dispute. The remaining daily capacity for those jurors with votes counted for this case is adjusted.

Experience Accumulation. Juror experience points are updated according to their participation. Juror's experience increases by 10 points after voting on a case. Finally, we note that in practice, jurors receive experience points through participating other types of cases in the Public Jury. Thus, to check how sensitive our simulation results are, we also run a scenario where we inflate experience points a juror received by voting in a case from 10 to 25. The results are summarized in Table 6 (Row 5).

F.2. Simulation results

The simulation results with standard errors are reported in Table F.2.

Table F.2 Policy Comparison (Full Results with Standard Errors)

	7-vote majority	16-vote majority	Dynamic (Experienced)	Dynamic (Mixed)	Inexperienced Jurors First	Increasing Enrollment
Seller win rate (%)	37.98	37.88	37.93	38.13	39.09	38.34
	(0.096)	(0.104)	(0.125)	(0.115)	(0.115)	(0.040)
Bias-S (%)	0.309	0.295	0.091	0.114	0.340	0.356
	(0.004)	(0.006)	(0.005)	(0.005)	(0.016)	(0.006)
Bias-B (%)	2.166	1.789	1.276	1.313	3.347	2.542
	(0.012)	(0.021)	(0.017)	(0.017)	(0.048)	(0.015)
Avg. # votes per case	8.34	19.0	8.33	8.35	8.38	8.35
	(0.003)	(0.008)	(0.005)	(0.004)	(0.004)	(0.002)
Vote exp point (Mean)	194149	150947	208448	193840	63187	182667
	(428)	(325)	(459)	(332)	(204)	(440)
Vote exp point (Median)	130714	88234	154870	129361	22524	112862
	(1098)	(663)	(954)	(1006)	(222)	(1065)
# Active Jurors at the end	42015	42220	42100	42159	42404	99743
	(25.8)	(28.3)	(46.9)	(32.6)	(39.6)	(54.3)
Avg. ending exp point	5278	6298	5332	5315	5111	2239
	(15.0)	(17.1)	(17.1)	(17.5)	(13.2)	(5.71)
# Juror with Exp Level ≥ 3	3105	3366	3136	3154	3267	3141
	(10.9)	(6.30)	(13.1)	(11.5)	(10.4)	(9.63)
# Juror with Exp Level ≥ 4	2335	2581	2360	2377	2452	2362
	(7.63)	(8.06)	(7.98)	(10.4)	(8.37)	(5.86)

Notes. The standard error is reported in the parenthesis.