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

Alan P. Kwan
Faculty of Business and Economics, University of Hong Kong, apkwan@hku.hk

S. Alex Yang
London Business School, sayang@london.edu

Angela Huyue Zhang
Faculty of Law, University of Hong Kong, angelaz@hku.hk

Disputes over transactions on two-sided platforms are common and usually arbitrated through platforms' customer service departments or third-party service providers. This paper studies *crowd-judging*, a novel crowd-sourcing mechanism whereby users (buyers and sellers) volunteer as jurors to decide disputes arising from the platform. Using a rich dataset from the dispute resolution center at Taobao, a leading Chinese e-commerce platform, we aim to understand this innovation and propose and analyze potential operational improvements, with a focus on in-group bias (buyer jurors favor the buyer, likewise for sellers). Platform users, especially sellers, share the perception that in-group bias is prevalent and systematically sways case outcomes as the majority of users on such platforms are buyers, undermining the legitimacy of crowd-judging. Our empirical findings suggest that such concern is not completely unfounded: on average, a seller juror is approximately 10% likelier (than a buyer juror) to vote for a seller. Such bias is aggravated among cases that are decided by a thin margin, and when jurors perceive that their in-group's interests are threatened. However, the bias diminishes as jurors gain experience: a user's bias reduces by 95% as their experience grows from zero to the sample-median level. Incorporating these findings and juror participation dynamics in a simulation study, the paper delivers three managerial insights. First, under the existing voting policy, in-group bias influences the outcomes of no more than 2% of cases. Second, simply increasing crowd size, either through a larger case panel or aggressively recruiting new jurors, may not be efficient in reducing the adverse effect of in-group bias. Finally, policies that allocate cases dynamically could simultaneously mitigate the impact of in-group bias and nurture a more sustainable juror pool.

Key words: crowd-sourcing, crowd-judging, platform governance, platform operations, two-sided marketplace, bias, experience, learning

1. Introduction

Conflict management and resolution is an essential function of two-sided online platforms. As Jeff Jordan, the former President of PayPal stated, "a big part of managing two-sided marketplaces involves managing tensions between the two, often opposing sides" (Jordan 2015). Traditionally, the governance of conflict resolution is centralized – platforms design rules for the participants,

apply and enforce these rules, and mediate and adjudicate conflicts whenever they arise. But such a model has faced two major challenges. First, the enormous amount of transactions facilitated by online platforms inevitably generates a high volume of disputes. For example, eBay estimated that it handled over 60 million cases each year (Del Duca et al. 2014). Platforms that solely rely on their internal customer service to resolve such conflicts may therefore incur heavy costs and experience significant delays, thereby damaging customer satisfaction and user loyalty (Rule and Nagarajan 2010). Second, a centralized model of governance is often criticized for the lack of procedural justice. This negative perception can call into question the legitimacy of the platform's decision-making which harms the relationship between the users and the platform (Van Loo 2016). This balance is particularly difficult to strike for a two-sided platform: pro-consumer outcomes come at the expense of sellers, while pro-seller outcomes disfavor consumers.

Faced with these challenges, large online platforms are experimenting with schemes to decentralize their governance by delegating some of their decision-making authority to users. One such innovation involves platforms crowd-sourcing their own users (as "crowd-jurors") to adjudicate disputes arising from the platform, a phenomenon which we call "crowd-judging". Crowd-judging was first trialled in 2008 by e-Bay India, who launched a community court which crowd-sourced buyers and sellers of eBay to adjudicate disputes regarding online feedback. Marketplatts, an online trading platform affiliated with eBay, launched GebruikersJury in 2011, a similar crowd-sourcing platform to resolve online disputes in the Netherlands. Alibaba followed suit and created a crowd-sourcing dispute resolution center in 2012 ("Taobao Public Jury"). Unlike eBay's community court, which only handles disputes of online feedback, Alibaba crowd-sources users to adjudicate a wide variety of disputes, and to vote on transactional rules and regulations for its e-commerce sites Taobao and T-mall. The success of Alibaba's crowd-sourcing mechanism has inspired other Chinese platforms, including second-hand marketplace Idle Fish, question and answer website Zhihu, food delivery company Meituan and restaurant rating firm Dianping to implement similar mechanisms.[1]

As crowd-judging is gaining popularity, it is important to understand whether this innovation alleviates concerns about the speed and quality of traditional centralized platform governance. In this paper, we provide the first empirical investigation into crowd-judging, using data from Taobao from June 2016 to February 2018. The dataset includes all transactional dispute cases decided on Taobao Public Jury, the votes cast by individual jurors (i.e., whether the juror votes in favor of the buyer or the seller), the time at which each vote is cast, and the final case outcome (majority ruling).

[1] In 2017, Idle Fish launched its crowd-judging tribunal to adjudicate disputes between buyers and sellers of second-hand goods. In 2019, Zhihu created a tribunal which crowd-sources eligible users to participate in content moderation. In 2020, Meituan Takeout and its affiliate Dianping.com, a restaurant rating platform, engaged their users to review disputes over feedback by applying similar crowd-judging mechanisms.

The dataset also includes information on the demographic characteristics of individual jurors such as age, gender, geographic location, jurors' status in the Taobao community (a registered buyer or seller), and their accumulated experience points rewarded for participation on Taobao Public Jury. Within the twenty-month time span, more than 155,500 jurors, among whom 80% were buyers, cast more than 6.2 million anonymous votes on more than 630,000 cases. User contribution is highly skewed: more than 90% of the total votes were cast by the top 10% of the jurors. Our analysis confirms that crowd-judging greatly improves the speed of dispute resolution: it takes 73 minutes for Taobao's crowd-judging center to collect sufficient votes to decide a median case, in contrast to what could normally take three to four business days when handled by Taobao's in-house customer service representatives (Alibaba Inc. 2016).

Though crowd-judging ensures rapid dispute resolution, critics have called into question the quality of such judgments. As in all other judgment mechanisms, crowd-judging could potentially suffer from two shortcomings: variability and bias. Although crowd-sourcing could effectively reduce the impact of individual judging variability on the aggregated outcome by virtue of the large number of individuals deciding a case (commonly known as the 'wisdom of the crowd'), such a phenomenon does not necessarily mitigate biases. Indeed, if a large fraction of jurors exhibit the same type of bias, a larger-sized crowd may amplify that bias, rather than dampen it.

In this paper, we focus on a type of bias commonly known as *in-group bias*, that is, buyer jurors have the tendency to rule in favor of buyers in disputes, and vice versa. This bias is regarded as a major quality concern of crowd-judging. Indeed, Taobao vendors often complain to the company and on social media that sellers are placed at a disadvantaged position in crowd-judging cases as the majority of crowd-jurors are buyers, who tend to decide cases in favour of buyers.

These sellers' worries are not unfounded. There is a rich literature that finds in-group bias emerges from artificially created social groups (Tajfel et al. 1971, 1979) and intrinsic social categorization such as ethnicity and race (Price and Wolfers 2010, Shayo and Zussman 2011, Anwar et al. 2012). Different theories, such as social identity and optimal distinctiveness, have been proposed to explain this phenomenon (Hewstone et al. 2002). Based on these prior research, it would not be surprising that there exists in-group bias based on users' seller/buyer status on the platform. However, it is not clear whether it is sufficiently severe so that it jeopardizes the legitimacy of crowd-judging. Motivated by the above practical and theoretical considerations, our first research question is: To what extent does in-group bias based on buyer/seller status exist among crowd jurors, and what factors mitigate or amplify this bias?

In addition, and more importantly from a managerial perspective, we note that in-group bias at individual level alone does not allow us to qualify the impact of such bias on case outcomes, which are decided through the majority ruling of a group of jurors. Thus, our second research question

is: how big is the impact of individual in-group bias on the case outcome in aggregate, and how could this impact be reduced through managerial interventions?

To answer our first research question, we exploit the quasi-random allocation of cases to individual jurors, and examine how seller and buyer jurors vote differently in the same case. In our preferred specification, which includes case fixed effects, we find that on average, a seller juror is approximately 10% likelier than a buyer juror to vote in favor of a seller in a dispute.[2]

Further analysis reveals three important factors that affect the magnitude of in-group bias. First, we find that jurors exhibit larger in-group bias in more ambiguous cases – as opposed to those clear-cut ones (cases involving clearly frivolous complaints or egregious behavior). Defining ambiguous cases as those decided by a thin margin by the first few votes, we find that the magnitude of in-group bias exhibited among the subsequent votes is 30% higher among those clear-cut cases.

Second, in-group bias amplifies as jurors perceive that their in-group's interest is being threatened. Utilizing the feedback of case outcomes the platform provided to jurors, we measure the level of perceived threat by how much more the majority rules in favor of a juror's outgroup than the juror themselves.[3] Controlling for case and juror fixed effects, a one-standard-deviation increase of this threat index could cause the in-group bias to increase by six percentage points. We also find evidence that in-group bias is likely to reside on both the buyer and seller side. Finally, this finding cautions platform operators that providing feedback to jurors may amplify bias.

Third, we identify a strong correlation between judging experience and in-group bias. Specifically, in-group bias among jurors with no experience can be as large as 23 percentage points, yet this gap reduces by 95% when the juror's experience is at the sample median level. This correlation persists as we include juror fixed effects. That is, as jurors gain more experience in deciding disputes, in-group bias reduces significantly. This behavior is most consistent with learning by doing and inter-group contact.

Combining the above empirical findings and a data-driven simulation that incorporates juror voting and participation behavior, a majority voting system, and different case allocation policies, this paper provides several managerial implications that could help improve the design and operation of crowd-judging on two-sided platforms. First, by varying users' buyer/seller status and using our estimate on individual in-group bias, we find that using the case allocation policy that Taobao

[2] While we could not completely rule out the possibility that jurors may skip certain cases the system distributed to them, our interview with Taobao representatives confirms that case skipping is rare. Further, we have conducted empirical checks, and found no evidence that jurors skip cases along the line of their buyer/seller status. We refer the readers to Appendix B for details.

[3] For example, out of the 10 cases a seller juror voted on a day, if the juror voted in favor of the seller side for six times, while the majority rules in favor of the seller for only two cases, then, upon receiving the final results, this juror may feel the interest of his in-group that is threatened more than if the majority rules in favor of the seller for six cases. See Section 5.3 for more details.

adopted, the in-group bias could influence no more than 2% of final case outcomes even under conservative assumptions. This result provides some assurance that in-group bias, one of the most prominent concerns around the use of crowd-judging, has a limited impact on case outcomes.

Second, using simulation, we estimate that, due to juror participation behavior,[4] increasing the required majority vote from seven (under the baseline policy) to 16 only results in a mild decline (4%-17% from the baseline scenario) in instances where case outcomes are affected by in-group bias, while the case resolution time could double. Similarly, we find that aggressively recruiting new users to serve as crowd-jurors could aggravate the impact of in-group bias because inexperienced jurors crowd out experienced ones. These findings alert managers that when bias is a major concern, simply increasing crowd size may not be a panacea.

Finally, based on our empirical findings, we design several policies that allocate cases dynamically among crowd-jurors. In one such policy, we first label a case as clear-cut or ambiguous based on whether the initial votes heavily favors one side. Subsequently, we allocate the ambiguous cases only to experienced jurors and the clear-cut cases only to inexperienced jurors. Our simulation result suggests that compared to the baseline policy, this policy could reduce the impact of in-group bias by 40% or more. The policy also improves the participation of inexperienced jurors, thus nurturing a more sustainable pool of crowd-jurors.

By providing the first empirical analysis of decentralized platform governance, the contribution of the paper is two-fold. First, we contribute to the academic literature by showing that significant in-group bias could arise due to the different economic roles of users in marketplaces, but that such bias can be mitigated as users gain more experience in judging. Second, we provide convincing evidence on the quality of crowd-judging and offer directions for further improvement. Taken together, the findings of our paper show the promise of using crowd-sourcing to fulfill yet another organizational function, that is, easing tension and resolving conflicts arising from two-sided platforms.

2. Literature

Our paper is related to several streams of literature. On application settings, it is related to three areas: crowd-sourcing, platform governance and operations, as well as judicial behavior and private ordering. Crowd-sourcing has received growing attention among scholars who recognized its potential as an innovative mechanism to leverage the collective intelligence of the crowd to deal with traditional organizational functions (Terwiesch and Xu 2008, Boudreau et al. 2011, Kremer et al. 2014, Huang et al. 2014). One line of research in this area has tried to assess the quality of

[4] We present empirical results on juror participation in Figure 2 and Appendix C. In general, we find that more experienced jurors are associated with shorter response time. In addition, despite the large number of jurors, the pool of experienced jurors is limited compared to the number of cases. Thus, more inexperienced jurors are involved as the panel size increases.

crowd-sourced decisions, mostly by comparing the decisions between the crowd and experts (e.g., Antweiler and Frank 2004, Larrick and Soll 2006, Budescu and Chen 2014, Mollick and Nanda 2015, Da and Huang 2020). In our setting, however, there is no "correct outcome" of the crowd's decision. Instead we focus on quantifying a specific form of quality: in-group bias, which has been attributed as a primary concern in practice and has the potential to systematically distort the outcomes. In this respect, our paper is closely related to Greenstein and Zhu (2012, 2018), who study bias in crowd-sourcing by examining the political ideology behind Wikipedia's editorial slant. Similar to their study which focused on the viewpoints of the crowd, we also examine a setting of "contested knowledge" since crowd-jurors possess wide discretion in adjudicating transactional disputes. Differing from their papers which examine bias exhibited by the crowd, we study both individual bias and its impact on case outcomes under majority ruling.

Our paper is also part of a growing body of research on platform governance. Thus far, scholars have explored issues such as who can access the platform (e.g., Boudreau 2010, Parker and Van Alstyne 2018) and how to create opportunities for inter-firm exchange and complementary innovation (Foerderer 2020). Our paper examines conflict resolution in two-sided markets, an important governance function that is under-explored in existing literature. An exception is Bakos and Dellarocas (2011), who examined conflict resolution on two-sided platforms and found that the traditional litigation-like mechanism for dispute resolution is more efficient than an online reputation system in inducing seller effort in a variety of settings. Unlike their paper, which focuses on platforms' centralized decision-making, we focus on an innovative decentralized dispute resolution mechanism. More recently, Papanastasiou et al. (2022) study how delegation could improve platform dispute resolution in the context of review blackmail, and Lee and Cui (2022) analytically examine the efficacy of crowd-sourcing based dispute resolution in the Gig Economy. Broadly speaking, our work is related to the growing literature on platform operations, where scholars have studied various issues such as pricing (Cachon et al. 2017, Bimpikis et al. 2019), ownership (Benjaafar et al. 2019), capacity control (Gurvich et al. 2019), agent retention (Musalem et al. 2019), review management (Xu et al. 2021), and its interaction with financing decisions (Chod et al. 2022, Cohen et al. 2020). This paper complements this literature by highlighting both the power of mechanism innovation in platform dispute resolution and the value of operational design in mitigating the impact of bias. To that extent, our study is also related to a stream of literature on identifying and mitigating discrimination in online markets (e.g., Edelman et al. 2017, Chan and Wang 2018, Cui et al. 2020, Mejia and Parker 2021).

Focusing on the application of crowd-sourcing in dispute resolution, our study also contributes to the legal literature on the behavior of judges and juries (Epstein et al. 2013, Rehavi and Starr 2014, Chen et al. 2016), and that on private ordering (Bernstein 1992, Greif 1993) and commercial

arbitration (Egan et al. 2018). In this paper, we not only quantify a potential concern of crowd-judging as a novel private dispute resolution mechanism, but also offer managerial insights on future improvement.

On underlying mechanisms, by focusing on in-group bias as the main quality measure, our paper is related to the rich literature of in-group bias, which can be at least traced back to the seminal work of Sherif et al. (1961) and Tajfel et al. (1971). Using experiments, these work shows that in-group bias is prevalent and arises even between artificially created social group. More recently, economists, legal and political scholars use naturally occurring data to quantify in-group bias in the context of judging, when in-group bias is based on traits such as ethnicity and race (e.g., Price and Wolfers 2010, Shayo and Zussman 2011, Anwar et al. 2012) and residence status (e.g., Klerman 2022). In addition to identifying in-group bias, this literature also studies the social and psychological motives behind this bias, its moderators, and remedies (Hewstone et al. 2002, Chen and Li 2009). Complementing this stream of research, we use natural occurring data to identify in-group bias based on a social categorization (buyer/seller) that is under-studied but highly relevant in business settings, and examined different moderators to in-group bias. These empirical findings, together with our simulation study, enable us to generate important managerial insights.

3. Empirical Setting and Hypothesis Development

This section first introduces the empirical setting of our study, and then develops four hypotheses.

3.1. Taobao Public Jury

In 2012, Alibaba launched the Taobao Public Jury, a crowd-sourced online dispute resolution center to resolve conflicts arising from its B2C e-commerce platforms (Taobao and T-mall). By the end of 2018, 16 million cases were distributed and completed on the platform. Over 4.3 million users had registered as crowd-jurors. Among them, over 1.7 million had cast votes; in total, they had cast over 100 million votes on these 16 million cases.[5]

Cases. The Public Jury acts as an internal service to different business lines within the company who have dispute resolution needs. These businesses decide what cases within their business functions could be brought up to the Public Jury. The cases resolved by crowd-jurors can be largely classified into two types: 1) cases that are initiated by the platform and focus on what the platform believes to be inappropriate behavior (such as aggressive language in reviews, exaggerating advertising); 2) transaction disputes initiated by unsatisfied buyers against sellers. For instance, suppose a buyer complains that the goods received do not match the description provided by the

[5] The institutional details of the Taobao dispute resolution center are based on our various interviews with employees at Taobao and Taobao's internal report. Data is available on Taobao's website: https://pan.taobao.com/.

seller and requests a refund. When the seller refuses this request, the buyer can escalate the case to the platform and then choose whether to use Taobao's customer service or the Public Jury to resolve their dispute. As this type of cases has direct monetary implications (when the case is ruled in favor of the buyer, the seller needs to offer a refund to the buyer), they are considered more controversial and the decisions also received most criticism. Our research focuses on this type of cases. Our dataset includes all transactional dispute cases between June 2016 and February 2018.

Jurors. Users who volunteer to act as crowd-jurors on the Public Jury need to meet several threshold requirements. First, the volunteer must be a registered user for over a year, which ensures that they have some familiarity with the type of disputes that could arise on e-commerce platforms. Second, the volunteer must have a good credit history on the platform. These criteria disqualify unreliable or dishonest candidates, and also deter malicious users from creating new accounts to vote and manipulate case outcomes. To address concerns about the volunteers' potential lack of experience, every newly qualified juror needs to go through some training and pass exams before they can start working on a particular category of cases. Taobao motivates crowd-jurors through a point and ranking system. Jurors earn some experience points for each case decided by them, and such reward does not depend on whether the juror's vote is in alignment with the majority decision. For example, a juror will collect 10 experience points after completing one transactional dispute case. The accumulated experience points will enhance the ranking of the jurors, whose rank could range from Level 1 (the least experienced) to Level 8 (the most experienced).[6]

Case Distribution. Taobao distributes cases to crowd-jurors by broadcasting them on the Public Jury system through a multi-layer randomization process: firstly, they wait until the number of outstanding cases of a specific type has reached a certain threshold, and then assign these cases randomly into different batches ("task packs"); secondly, different task packs are sent randomly to a subset of registered jurors, and the sequence of cases in the same task pack is also randomized for different jurors receiving it. This randomization process is to minimize the possibility of users purposely coordinating with each other in an attempt to manipulate the outcome of certain cases. The jurors who are logged on to the crowd-sourcing center will be able to see whether there are available tasks for a specific case type (e.g., transactional dispute). The available cases are presented to the jurors sequentially, and jurors will need to finish reviewing the cases they are working on before moving on to the next. After reviewing the evidence of one case, jurors also have the option to not cast a vote on the case and move to the next one. However, based on Taobao's internal estimate, instances of jurors skipping cases are rare.[7]

[6] See Appendix A.3 for the mapping between experience points and the Public Juror Experience Level.

[7] Based on our interviews with Taobao employees managing Public Jury, normal jurors have no direct incentive to skip cases as they are rewarded with experience points regardless of whether their vote is consistent with the final

Judging. When judging a case, jurors are able to review the evidence submitted by the buyer and seller in a dispute, but the identities and other sensitive information (e.g., address) of the parties in dispute are kept anonymous for privacy reasons.[8] Jurors cast votes anonymously and independently. They only receive the vote tally until the next time when they log on the system if the case has been concluded by then.

Cases are determined under a simple majority ruling. The number of votes needed to determine the outcome of a case has been changed twice in our sample period (June 2016 – February 2018). In the beginning, the party that first received a majority vote of 16 would have won the case, thus the panel size could range from 16 to 31. In August 2017, Taobao reduced the majority vote requirement from 16 to 7, thus narrowing the range to fall between 7 to 13 (the "7-votes period"). However, it reverted back to the majority vote of 16 in January 2018.[9] Once a case has accumulated sufficient votes to form the required majority, it is removed from all relevant jurors' case list. If the majority of the jurors vote in favor of the buyers, then Taobao will enforce the decisions, for example, by freezing the payment in dispute or taking money from the store deposits of the sellers. After being decided by the crowd on Public Jury, a case can be appealed. According to Taobao's internal estimate, only a small fraction of such cases are appealed to customer service.[10] Finally, using crowd-judging imposes no other penalty or cost to either party in a dispute than having the case adjudicated by a different channel (e.g., through Taobao customer representatives). For example, if the decision by the crowd is unfavorable to the seller, it will not affect their rating or reputation on the e-commerce platform differently from if the case is ruled against the seller by a Taobao customer representative.

3.2. Hypotheses Development

Connecting the above empirical setting to the extant literature, we develop four hypotheses centered around in-group bias along jurors' buyer/seller status in crowd-judging.

Existence of in-group bias. According to the in-group bias literature, two possible channels may contribute to the existence of this bias in our setting. First, as amateurs, crowd-jurors are likely to be influenced by their personal experience as a buyer or seller, and thus have more

majority rule. Taobao also believes that the number of malicious jurors hired by one party in dispute to manipulate case outcome is kept at a very low level, if any at all, under the sophisticated case distribution policy that they adopt. In addition, we conducted empirical checks, and found no evidence that jurors skip cases along the line of their buyer/seller status (Appendix B). This assures us that even case if skipping behavior exist along other dimensions, they are likely to be orthogonal to our identification of in-group bias, which is based on jurors' buyer/seller status.

[8] When certain sensitive information is essential for judging a case, this case will be adjudicated by Taobao employees.

[9] At the same time of the voting rule changes, the internal client also adjusted the number of cases distributed around the same time as the required number of votes was changed. We discuss these changes in Appendix C.3.

[10] In our dataset, out of more than 600,000 cases, only less than 100 were marked as having been brought to appeal.

sympathy for their in-group and hold more negative feelings about the out-group. This could lead them to be more inclined to vote in favor of their in-group party. This form of discrimination tends to be implicit, and the jurors themselves may be unaware of it (Greenwald and Banaji 1995, Gawronski and Bodenhausen 2006). The other potential channel is explicit bias, that is, preference and attitudes that people are consciously aware of (Carter and Murphy 2015, Sommers and Norton 2006). For instance, some disgruntled consumers join Taobao's crowd-judging platform because they see it as an opportunity to seek revenge against sellers (even if they are unaware of who the specific seller in the crowd-judging case is). Based on the above reasoning, we form our first hypothesis.

HYPOTHESIS 1. *Everything else being equal, seller jurors are more likely to vote in favor of the seller party in disputes than buyer jurors are.*

In-group bias and case ambiguity. Case characteristics, and specifically, whether a case is ambiguous could also affect the magnitude of in-group bias.[11] One reason could be that jurors are more likely to resort to their intuition when making a judgment under uncertainty, which could be affected by unconscious bias (Tversky and Kahneman 1974). Moreover, uncertain cases provide judges with the opportunity to exercise more discretion, allowing the personal preference of jurors to play a role in affecting the final outcome (Posner 2010). Both theories lead to the following hypothesis.

HYPOTHESIS 2. *Crowd-jurors exhibit more in-group bias when facing more ambiguous cases.*

In-group bias and perceived threat. Several theories of in-group bias view "threat" as a "central explanatory concept", and predict that perceived threat heightens in-group bias (Hewstone et al. 2002). Empirically, Quillian (1995) find that in-group bias and prejudice are associated with certain social group's perceived threat on the inter-group competition over scarce resource. In our setting, jurors could see the case outcome (as determined by the majority ruling) once the case is concluded. Thus, when crowd-jurors see that their in-group side loses on cases that they believe the in-group side should win, these jurors may form a perception that their in-group members' economic interests are being threatened. Subsequently, they may vote in a more biased way.

HYPOTHESIS 3. *Crowd-jurors exhibit a higher degree of in-group bias when they perceive that their in-group side's economic interests are being threatened.*

[11] Case ambiguity could be caused by two scenarios. First, the parties in dispute do not provide clear evidence for crowd-jurors to make informed decisions; second, that both sides are somewhat at fault based on evidence presented, so how to judge this case is contested knowledge As discussed later, as we do not have data on case characteristics, we use initial votes to construct a measure for case ambiguity.

In-group bias and judging experience. Our final hypothesis is on the relationship between jurors' judging experience and in-group bias. Judging experience could affect in-group bias through two channels: inter-group contact and learning by doing. First, in the in-group bias literature, inter-group contract (interactions between membership of different groups) has been identified as an effective approach in reducing in-group bias (Hewstone et al. 2002, Dovidio et al. 2017). In our setting, when making judging decisions, jurors need to review evidence provided by both sides in the dispute. This experience provides them with opportunities to understand and appreciate the perspective of their out-groups. In this sense, judging experience serves as a form of inter-group contact. Thus, more judging experience is associated with more inter-group contact, and thus less in-group bias.

Judging experience is also related to learning by doing. Prior studies have established learning by doing as an effective means to improve quality and speed in contexts such as manufacturing (Shafer et al. 2001, Levitt et al. 2013), service delivery (e.g., Huckman and Pisano 2006, Ramdas et al. 2018) and development (Boh et al. 2007). In our setting, as we argued when developing Hypothesis 1, one source of in-group bias originates from the fact as jurors are mostly amateurs with little judging experience, they are more likely to draw on their own experience as a buyer or seller when making judgments. However, as they vote on more cases and gain more judging experience, they learn to better assume the role of a judge and thus make less discriminatory decisions.[12]

HYPOTHESIS 4. *Crowd-jurors exhibit less in-group bias as they gain more judging experience.*

Despite the above theories that support this hypothesis, we also note two caveats. First, in a setting similar to ours, Shayo and Zussman (2011) do not find a positive correlation between judge experience and in-group bias that is related to ethnicity. Second, the prior literature on learning by doing largely focuses on users' professional settings where agents receive monetary compensations. In contrast, our setting involves crowd-jurors acting as volunteers without any monetary rewards.

4. Data and Summary Statistics

Our dataset includes all 618,329 transnational dispute cases judged on the crowd-judging platform from June 1, 2016 to February 28, 2018. Over this 20-month period, in total 155,520 crowd-jurors cast 6,242,315 votes on those cases.[13]

[12] Another form of learning is observational learning (Cai et al. 2009, Cui et al. 2019, Aksin et al. 2021). In our setting, because jurors do not communicate with each other while judging a case and their identities are kept anonymous when they vote, there is little interaction among the crowd-jurors. Moreover, the crowd-jurors receive no reward for making a decision in line with the majority, nor any punishment if its decision is in the minority. As such, they face little social pressure to conform to others. In such a scenario, social learning, if it exists, is rather indirect and the effect would be attenuated in our setting.

[13] As we have only obtained data on a specific type of case (transactional dispute) over this 20-month period, the number of participating jurors in our sample is significantly lower than the total number of jurors who ever registered or voted in Public Jury.

Table 1 Summary Statistics

	N	Mean	Std. Dev.	25th	Median	75th
% of vote in favor of seller (by case)	618,329	40.62	32.6647	12.5	33.33333	70
Number of cases decided (by juror)	155,520	40.1383	554.25527	1	3	8
Experience points (by juror)	155,520	2,810	21,100	60	141.25	500
Experience points (by vote)	6,242,215	223,082	247,329	10,547	133,547	362,127

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.

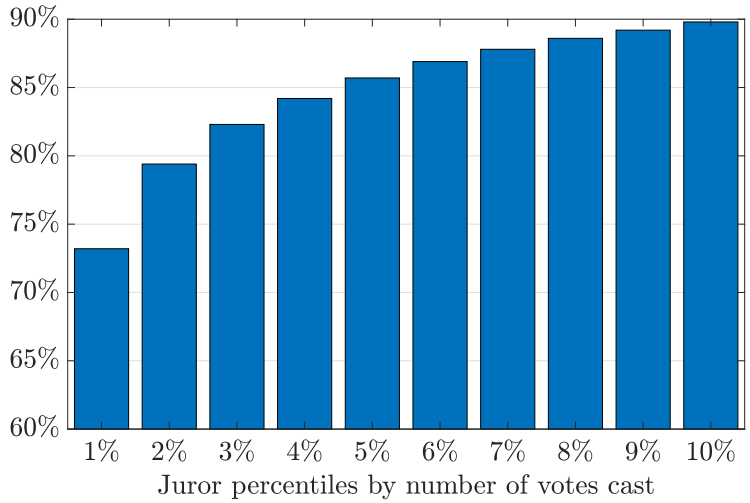
The summary statistics are presented in Table 1. Row 1 reports the distribution of votes in favor of seller at the case level. As shown, the median case settles in favor of the buyer, with one third of votes going to the seller. At juror level (Row 2), 45% of the 155,520 jurors in our sample are female. In terms of user participation, jurors on average completed 40 cases during the sample period, with the median juror completing three. This reveals that jurors' participation on the platform is skewed. Two pieces of evidence further support this pattern of skewed participation. First, as shown in Figure 1, in the dataset, more than 90% of the votes are cast by the top 10% of most active jurors, with the top one percent casting more than 70% of the votes. Second, as shown in Rows 3 and 4 in Table 1, juror experience points in the Public Jury system are also highly skewed, at both user and vote level. As shown, the median juror's average experience point in the sample is 150. In contrast, the median vote is cast by a juror with 133,547 points, who has voted on more than 10,000 cases on the platform.

This skewed contribution is consistent with the findings in previous literature that the participation in crowd-sourcing platforms is highly concentrated, where a minority of users make the majority of contributions (Franke and Von Hippel 2003, Sauermann and Franzoni 2015). Further, we note that given this skewed contribution, the potential bias of the most active and experienced jurors is expected to have a significant impact on the final case outcome.

Further, we also observe that juror experience are correlated with their response speed. Figure 2 plots the average juror experience for the kth vote in each case. For example, among all cases decided during the 7-votes-majority period (Figure 2(a)), the average experience point for the jurors that cast the first vote in each case is 242,515, while it drops to 118,350 on the 7th votes. It then further decreases to 76,295 on the 13th vote, when needed. Figure 2(b) reveals the same pattern during the 16-majority-vote period. As shown in Section 6, such voting pattern, together with the skewed contribution, has important implications on the design of policies for crowd-judging.

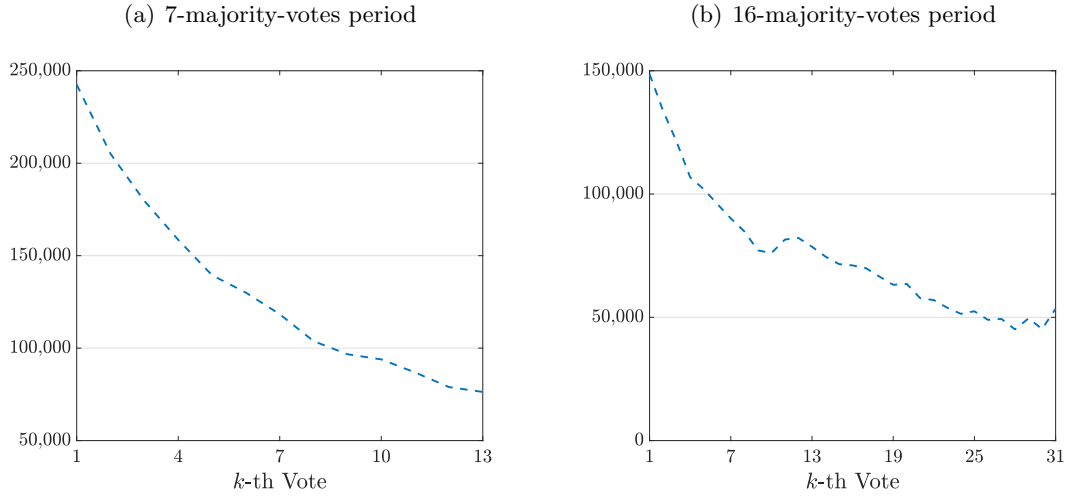
Finally, our data confirms that crowd-judging resolves cases at an impressive speed. For example during the 7-votes majority period, the median case resolution time (from the time that a case

Figure 1 Fraction of Votes Contributed by Top 10% of Jurors



Notes. The x-axis represents the ith-percentile ($i = 1, \ldots 10$) of jurors ranked by the number of votes they cast over this period. The y-axis represents the cumulative fraction of votes cast by jurors above or equal to this percentile as a fraction of total votes.

Figure 2 Average Experience Point by Vote Ranks



Notes. The x-axis represents the kth-vote ($k = 1, \ldots 13$ for the 7-majority-votes period, and $k = 1, \ldots 31$ for the 16-majority votes period) as ranked by when the vote is cast in each case. The y-axis represents the average experience points for the jurors that cast the kth vote in each case at the time of voting.

is submitted to the Public Jury to when sufficient votes are collected to decide the case) is 73 minutes, and more than 75% of the cases are decided within 18 hours.[14] Such resolution speed is

[14] The median case resolution time during the 16-votes majority period is 22 hours. Further, if we focus on the time it takes a case to collect all votes (the time elapsed from the first vote to the last one), the median is 31 minutes

significantly quicker than that of customer service, which, according to Taobao, usually takes three to four days (Alibaba Inc. 2016).

5. Empirical Findings

This section starts by establishing the existence of in-group bias, and then presents empirical findings on how three case- and juror-level factors that affect the magnitude of in-group bias.

5.1. Existence of In-group Bias

To test for the existence of in-group bias, we conduct an analysis at the individual vote level.[15] We estimate an ordinary least squares regression of the following specification:

$$VoteSeller_{ijt} \times 100 = \beta \times Seller_j + X'_{jt}\gamma + \eta_t + \delta_i + \epsilon_{ijt}, \tag{1}$$

in which $VoteSeller_{ijt}$ is the binary variable indicating whether the vote cast for case i by juror j at time t is in favor of the seller ($VoteSeller_{ijt} = 1$) or the buyer ($VoteSeller_{ijt} = 0$). We multiply this number by 100 so that the coefficients can be directly interpreted as percentage points. $Seller_j$ is the dummy for whether juror j is registered as a seller (1) or a buyer (0), and X_{jt} is a vector of juror characteristics including their age, gender, and registered province, and experience point on the crowd-judging platform at the time of the job. δ_i is the case fixed effects, and ϵ_{ijt} is an error term double clustered at juror and case level.[16]

The coefficient of interest is β: if in-group bias indeed exists, we expect β to be positive and statistically significant. For a large portion of our empirical tests, we employ case fixed effects to control for confounders. For example, although the case distribution process is effectively random, seller jurors and buyer jurors may, for various reasons, skip cases with different characteristics. In this case, without case fixed effects, the estimated β may capture these case characteristics instead of in-group bias. However, with case fixed effects, we could control any unobserved case heterogeneity and thus render the estimation "within-case", mitigating selection concerns. Crucially, while we do not know whether the true outcome for any given case ought to be in favor of a buyer or a seller, *within* a case, in-group bias can still be identified if sellers and buyers are likelier to vote for their respective groups. Further, the need to compare jurors within the case necessitates our usage of OLS rather than a logistic regression, because logistic regressions suffer from the "incidental

during the 7-votes-majority period, and 175 minutes during the 16-votes-majority period. See Appendix C.1 for more details.

[15] We perform the analysis at the individual vote rather than case level because it enables us to explore variation between users in the same case or across users over time. Both of these dimensions of variation are removed at the case level.

[16] Conservatively, we present double-clustered standard errors. We perform additional analyses to show that our results remain unchanged under other clustering methods. See Appendix D.1.2 for details.

parameters" problem (Greene 2004). Another benefit of OLS is that marginal coefficients are easier to interpret as linear probabilities.[17]

Table 2 Existence of In-group Bias

Dependent Variable: VoteSeller ×100								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Seller	4.68***	4.07***	4.38***	3.87***	6.38***	5.18***	3.54***	3.71***
	(1.28)	(1.11)	(1.45)	(1.26)	(0.934)	(0.794)	(1.23)	(0.843)
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓	✓	✓
Case FE		✓		✓		✓	✓	✓
Sample Period	Full	Full	7-votes	7-votes	16-votes	16-votes	7-votes	16-votes
Votes	All	All	All	All	All	All	First 7	First 7
Observations	6,242,315	6,240,035	5,088,674	5,088,674	1,153,641	1,153,641	3,919,349	408,954
R-squared	0.0654	0.38019	0.04879	0.3764	0.14091	0.39704	0.4348	0.5072

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 2 contains the resulting coefficient estimates for the above model. We present eight specifications to demonstrate the robustness of our main finding under different model specifications (with or without case fixed effects) and different samples. Column (1) displays the relationship between the probability a juror will vote for the seller and their status as a user of the platform with juror-level controls (e.g., gender, age) and month fixed effects. If the crowd-judging program or the composition of jurors undergoes variation over time, the month fixed effect accounts for some of this variation. In this specification, a seller is 4.68 percentage points likelier to vote in favor of the seller, which is indicative of in-group bias. In Column (2), we include case fixed effects, which increases the R-sq of the regression drastically. The coefficient for in-group bias, on the other hand, remains resilient, only dropping slightly to 4.07%. This magnitude is economically significant: Considering that just under 40% of votes are in favor of the seller, this bias of 4.07% can be interpreted as sellers being approximately 10% ($= \frac{4.07\%}{40\%}$) likelier to vote for other sellers.

Columns (3)–(6) perform the sub-sample analysis in our two different sub-periods, one where the case is determined with a majority of seven votes ("7-votes period", Columns 3–4) and the other 16 votes ("16-votes period", Columns 5–6). In general, we find that the in-group bias is robust in all sub-samples. In addition, we find that in-group bias is marginally larger during the 16-votes

[17] As a robustness test, Appendix D.1.1 show that the estimate of in-group bias is also statistically significant under a logistic regression specification.

period than the 7-votes one, although the difference is not statistically significant. We conjecture that there are two possible reasons. First, it is possible that individual juror's behavior changes in response to the change of judging rules (from a 7-votes majority to a 16-votes majority). For example, an individual juror may perceive his/her role in a larger panel as less important, and may thus make decisions less carefully. We call this the "behavior effect". An alternative explanation is that as we observed in the Section 4, during the 16-votes period, more inexperienced jurors participate in judging during the 16-votes period than during the 7-votes one. This composition change could lower judging quality (the "composition effect"). To shed light on the importance of these two effects, we repeat the above sub-period analysis, but only use the first seven votes in each case, mitigating the composition effect. The results are presented in Columns (7)-(8). As shown, the in-group bias between the two sub-sample periods largely diminishes when examining the first seven jurors. This suggests that juror composition plays a large role in determining the magnitude of (average) in-group bias.[18] Finally, we augment Eq. (1) with interaction between three juror characteristics (age, gender, and one measure related to jurors' geographic region) and the *Seller* indicator. We find that both the magnitude and statistical significance of the in-group bias estimates remain largely unchanged, suggesting that in-group bias is prevalent and not driven by a particular sub-group of jurors. See Appendix D.1.3 for details.

5.2. In-group Bias and Case Ambiguity

We test Hypothesis 2 using the following specification.

$$VoteSeller_{ijt} \times 100 = \beta \times Seller_j + \phi \times Seller_j \times Ambiguity_i + X'_{jt}\gamma + \eta_t + \delta_i + \theta_j + \epsilon_{ijt}, \qquad (2)$$

Compared to our base specification (Eq. (1)), Eq. (2) includes two additional terms. First, the interaction term $Seller_j \times Ambiguity_i$, where $Ambiguity_i$ is an indicator of whether case i is considered (relatively) ambiguous, which we detail later;[19] This interaction term is our main variable of interest. Based on the above hypothesis, we expect the coefficient ϕ to be positive and statistically significant. Second, we add the user fixed effect θ_j for juror j. Similar to the case fixed effects, user fixed effects control for unobservable juror characteristics, alleviating concerns such as the estimate is driven by the personal preference of jurors rather than case ambiguity.

Next, as we do not have an external measure for case ambiguity, we construct the dummy variable $Ambiguity_i$ based on jurors' votes in the case. Intuitively, if the numbers of votes in favor of seller

[18] Appendix C examines how the panel size change affects of juror composition in more details. In addition, the comparison between Columns (6) and (8) also suggests that jurors with more experience, who tend to respond faster, are less biased than inexperienced jurors. This result is also consistent with our finding in Section 5.4 that jurors' in-group bias decreases as they gain more judging experience.

[19] The term *Ambiguity* itself is subsumed into the case fixed effects.

and buyer are similar (e.g., a 6-7 split), then the case is more likely to be ambiguous; otherwise (e.g., a 7-0 split), the case is more clear-cut. However, if we use all jurors' votes to construct the ambiguity measure, our measure of $Ambiguity_i$ would have a mechanical correlation with our outcome variable. To avoid this concern, we split the votes in each case into two parts: we use the first several votes in the case to construct $Ambiguity_i$, and then use the subsequent votes as observations in estimation. This way, we separate the votes used in estimating in-group bias from those involved in the determination of whether a case is ambiguous. In addition, due to the simple majority rule used to determine the case outcome, the total number of votes in a case could also be endogenous to case ambiguity. For instance, an ambiguous case may require a total 13 votes to form a majority of seven, while a landslide case could be determined in seven votes. To address this issue, we focus only on the votes required to reach a decision (e.g., the first seven votes in cases during the 7-votes period). For example, if we use 3 votes to designate ambiguity and 7 are required to form a majority, we only consider the impact of ambiguity on votes 4-7.

Table 3 In-group Bias and Case Ambiguity

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	3.34***		3.85***		4.89***	
	(0.852)		(0.570)		(0.668)	
Seller × Ambiguous	1.17***	0.797**	1.09***	0.956**	2.17***	1.79***
	(0.358)	(0.391)	(0.422)	(.0450)	(0.499)	(0.500)
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE		✓		✓		✓
Sample Period	7-votes	7-votes	7-votes	7-votes	16-votes	16-votes
Votes for Ambiguity	1–3	1–3	1–5	1–5	1–5	1–5
Vote Margin	1	1	1	1	1	1
Votes as DV	4–7	4–7	6–7	6–7	6–16	6–16
Observations	2,248,368	2,248,368	1,124,184	1,124,184	618,607	618,607
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.

We present our results in Table 3. In Columns (1)–(2), which focus on the 7-vote period subsample, for case i, we use the first three votes to construct $Ambiguity_i$: if there is a 1-2 split out of the first three votes, that is, one or two jurors in the first three vote in favor of the seller (a vote margin of 1), we define $Ambiguity_i = 1$. Otherwise, $Ambiguity_i = 0$. We then use the 4th–7th

votes in the case as dependent variables for our regression. In this case, we find that controlling for case fixed effects and other controls, the in-group bias appears to be substantially larger when a case is ambiguous. The average in-group bias is 3.34% in non-ambiguous cases, but is 33% larger in ambiguous cases. Column (2) includes user fixed effect, thus the *Seller* term is subsumed as it is fixed for each juror. The estimate says that the in-group bias is larger when the same juror faces an ambiguous case (relative to an unambiguous one), suggesting that our results are not generally explained by biased jurors matching with ambiguous or unambiguous cases. Columns (3)–(4) apply a different standard to classify ambiguous and unambiguous cases: $Ambiguity_i = 1$ if there is a 2-3 split out of the first five votes. The remainder requisite two are used as observations. Similarly, the results for the 16-votes period are presented in Columns (5)–(6) in Table 3. As shown, jurors continue to exhibit greater in-group bias when facing ambiguous cases. Additional robustness checks are presented in Appendix D.2, confirming our findings are robust with respect to alternative definitions of case ambiguity and controlling for juror experience.

We note that the above result also supports the interpretation that the difference in voting patterns between buyers and sellers is likely a form of bias rather than buyer/seller jurors selecting different types of cases. Specifically, for the observed buyer/seller difference to be caused by a form of selection bias, the most likely scenario is that jurors simply skip cases that their in-group side is likely to lose. In that case, as it should be more difficult to predict which side will win in a more ambiguous case, one would expect the buyer/seller difference to be smaller in ambiguous cases. This conjecture, however, runs the opposite to our empirical results. Another possibility is that jurors strategically choose cases that they believe are ambiguous and thus can be more easily influenced by their votes. However, we note that this interpretation still relies on the fact that jurors exhibit in-group bias. Further, as we explain in Section 5.4, this type of case selection is also unlikely to be prevalent.

5.3. In-group Bias and Perceived Threat

To test Hypothesis 3, we exploit the information that the Public Jury shared with jurors regarding the outcome on the previous cases these jurors have voted on. Specifically, when a returning juror enters the Public Jury system, they will be shown the outcomes, as well as their own votes, for all the concluded cases that they voted in during their last active day on the platform. Intuitively, a juror may feel their in-group side is at a more vulnerable position when observing his in-group side losing more cases than what the juror believe they should, as reflected by the juror's own voting patterns. To capture this intuition, we construct a juror-day level measure of perceived threat, $NetOut_{jt}$ as follows.

$$NetOut_{jt} = \frac{N^{case}_{j,t-1} - N^{vote}_{j,t-1}}{N_{j,t-1}}, \tag{3}$$

where $N_{j,t-1}$ is the total number of cases juror j participated in 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 example, consider juror j, who is a buyer. On this juror's last active day before time t, he participated in 10 cases ($N_{j,t-1} = 10$). Out of this 10 cases, the seller side (the focal juror's outgroup) won six times under majority ruling ($N^{case}_{j,t-1} = 6$), yet juror j voted in favor of the seller 4 times ($N^{vote}_{j,t-1} = 4$). Then we have $NetOut_{jt} = \frac{6-4}{10} = 0.2$. Essentially, this measure is positive when the majority favors the juror's out-group than the juror's own vote, and negative otherwise. Thus, the greater is $NetOut_{jt}$, the juror feels the juror's in-group is under greater threat. In response, the juror shall exhibit stronger in-group bias under subsequent voting. In contrast, if jurors may view this feedback as an opportunity to learn from other jurors, then by observing that the majority rules against theirs, the juror may become more aware of their in-group bias, thus vote in a less biased way.

To distinguish Hypothesis 3 from the above possibility, we employ the following specification:

$$\begin{aligned} VoteSeller_{ijt} \times 100 = \alpha \times Seller_j + \beta \times NetOut_{jt} + \phi \times Seller_j \times NetOut_{jt} \\ + X'_{jt}\gamma + \eta_t + \delta_i + \theta_j + \epsilon_{ijt}. \end{aligned} \tag{4}$$

The coefficient of interest is ϕ: a positive ϕ is consistent with the hypothesis that a higher level of perceived threat aggravates in-group bias, while a negative one suggests feedback on past bias could mitigate in-group bias.

The results are presented in Columns (1)-(2) in Table 4. As shown, ϕ is positive and statistically significant, consistent with the threat hypothesis. The impact is also economically large: a one standard deviation change in $NetOut$ ($\sim$20%) leads to a $31.29 \times 20\% = 6.2\%$ change in in-group bias, which is comparable to the average in-group bias as estimated in Table 2. Columns (3)-(4) show similar impacts are observed in both sub-periods. Additionally, Columns (5)-(6) estimate the impact of threat on buyer and seller separately. Among buyer jurors, a higher threat level reduces their likelihood to vote in favor of sellers (their out-group). Symmetrically, the tendency for seller jurors to vote in favor of their in-group increases significantly as they perceive higher levels of threat. Combined, these results suggest that although we cannot accurately estimate which side the in-group bias resides at, both sides are likely to exhibit a certain degree of bias.[20]

Finally, we note that as the feedback from the platform is the only channel through which jurors could form such this perceived threat, our result also points out that jurors do respond to the feedback the platform provides. While the conventional wisdom is that properly presented feedback could help improve agents' behavior, our result reveals that feedback could also have an unintended negative consequence as it could worsen decision-makers' biases.

[20] We also conducted a number of robustness tests, which are presented in Appendix D.3. For example, we decompose the measure of perceived threat $NetOut_{jt}$ into a positive contributor (the majority rules in favor of the out-group, but the juror voted for the in-group) and a negative one (the juror voted in favor of the out-group, yet the majority rules for the in-group). The result shows that jurors react to both components significantly, and with similar magnitude.

Table 4 In-group Bias and Perceived Threat

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	1.57***					
	(0.695)					
NetOut	−50.44***	−16.89***	−12.45***	−16.64***	−17.14***	10.86***
	(1.64)	(0.830)	(0.810)	(1.52)	(0.813)	(1.96)
Seller × NetOut	101.7***	31.29***	24.30***	25.63***		
	(3.51)	(3.77)	(3.25)	(2.56)		
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE		✓	✓	✓	✓	✓
Sample Period	Full	Full	7-votes	16-votes	Full	Full
Juror Sample	Both	Both	Both	Both	Buyer	Seller
Observations	5,488,367	5,488,367	4,490,855	997,512	4,690,441	797,926
R-squared	0.449	0.489	0.487	0.524	0.502	0.737

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.

5.4. In-group Bias and Judging Experience

We investigate this relationship between in-group bias and judging experience (Hypothesis 4) using the following specification:

$$\begin{aligned} VoteSeller_{ijt} \times 100 = \alpha \times Seller_j + \beta \times Exp_{ijt} + \phi \times Seller_j \times Exp_{ijt} \\ + X'_{jt}\gamma + \eta_t + \delta_i + \theta_j + \epsilon_{ijt}, \end{aligned} \tag{5}$$

where Exp_{ijt} is a measure for juror j's judging experience on the Public Jury at time t when casting vote on case i. In the main body of the paper, we measure experience (Exp_{ijt}) using $LogExp$, which is defined as the natural logarithm of one plus the juror's experience point at the beginning of the day when judging case i (in our dataset, juror experience points are updated daily). The log transformation is to reduce the skewness of juror experience points as shown in Table 1. The coefficient of interest is ϕ: when in-group bias is mitigated by juror experience, we would expect this coefficient to be negative. We employ different variants for this measure and the results remain robust. See Appendix D.4 for more detail.

In Table 5, Columns (1) and (2) present the regression results for the full sample, one without user fixed effects and one with. As shown in Column (1), seller jurors with zero experience points are 23% likelier to vote for a seller than zero-experienced buyer jurors. Further, the coefficient corresponding to the interaction term ($\phi = 1.86$) is not only statistically significant, but also economically meaningful. For example, the in-group bias among jurors with median level of experience

Table 5 In-group Bias and Juror Experience

Dependent Variable: VoteSeller ×100						
	(1)	(2)	(3)	(4)	(5)	(6)
Seller	23.37***		21.37***			
	(3.04)		(3.31)			
LogExp	−0.227	−0.460*	−0.253	−1.43***	−1.37***	−1.59***
	(0.173)	(0.275)	(0.196)	(0.292)	(0.394)	(0.2780)
Seller × LogExp	−1.86***	−2.20**	−1.71***	−1.45	−2.17*	−3.14**
	(0.390)	(1.10)	(0.424)	(0.989)	(1.12)	(1.46)
Controls	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Case FE	✓	✓	✓	✓	✓	✓
User FE		✓		✓	✓	✓
Sample	Full	Full	7-votes	7-votes	7-votes	7-votes
User Active Days	≥ 1	≥ 1	≥ 1	≥ 1	≥ 25	≥ 100
Observations	6,242,315	6,242,315	5,088,674	5,088,674	3,593,477	3,106,809
R-squared	0.381	0.476	0.377	0.474	0.650	0.533

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.

(experience points = 133,547) is only 1.3% ($= 23.25\% - \log(133,547) \times 1.86\%$). This translates to a 95% ($\approx \frac{23.2\%-1.3\%}{23.2\%}$) reduction compared to jurors with no experience. Nevertheless, the result does not specify whether the difference is due to the growth of each individual juror or variations across different jurors. One argument favoring the latter possibility is that less biased jurors, who are more likely to be motivated by real interests in resolving disputes, rather than personal revenge, are more likely to stay longer and judge more on the platform. In this case, certain unobservable juror quality could be a confounding factor, and may thus bias our estimation of ϕ. To control for this effect, Column (2) includes juror fixed effects. This allows us to capture how an individual juror's experience growth affects the juror's in-group bias. As the coefficient of $Seller \times Exp.$ remains negative and statistically significant, it directly supports our hypothesis that as a juror gains more experience over time, they exhibit less in-group bias. Columns (3)–(4) repeat the above analysis for the 7-votes sub-sample period. As shown, while the interaction term in Column (4) is not statistically significant, the magnitude is only marginally smaller. One possibility is that many jurors who voted on just a few days may contribute statistical noise to our sample. To alleviate this concern, we focus on users who participate for at least 25 or 100 days. The results in Columns (5)–(6) show that such sub-setting strongly recovers our results, suggesting that learning is observed at least among jurors who participate over a minimal number of occasions. In Appendix D.4, we continue to observe strong correlation between experience and in-group bias in the 16-votes sub-sample.

In summary, our empirical results provide consistent evidence that judging experience mitigate in-group bias, consistent with both hypothesized channels, namely, inter-group contact and learning by doing. These two channels could also help reconcile our findings with Shayo and Zussman (2011), who do not find a positive correlation between judging experience and ethnicity-based in-group bias. First, in the setting of Shayo and Zussman (2011), where the in-group bias is based on ethnicity, judges may already have a lot of experience interacting with different ethnic groups in their everyday life. Thus, judging experience may not attribute much to inter-group contact. In contrast, jurors in our setting have relatively less exposure to their out-group's perspective in dispute. Thus, the inter-group contact associated with judging experience could play a more significant role. Second, Shayo and Zussman (2011) examines the behavior of professional judges, who have received extensive legal training. Thus, learning by doing (especially considering that they are focusing on relatively routine cases) may not be as significant as in our setting, where most, if not all, crowd-jurors have no prior experience in adjudication.

Further, the within-juror experience result helps us alleviate the concern that our findings are (partly) driven by case selection. Recall that in Section 5.2, to reconcile the finding that jurors exhibit more in-group bias among ambiguous cases with case selection, one has to assume that jurors purposely choose ambiguous cases so they could influence the result in favor of their peers. While this type of case selection does not undermine our in-group bias interpretation, it makes the inference less clear. However, our result on experience suggests that this type of case selection is also unlikely. Assume the greater in-group bias among ambiguous cases is indeed due to jurors choosing cases that are easier to manipulate, then one would expect that as jurors become more experienced, they should be better at identifying cases that are ambiguous. Hence, we should expect in-group bias not to decrease with experience. Yet, we see the opposite.

Finally, we note that the above result also leads to an interesting direction on the design of crowd-judging policies: on the one hand, allocating more cases to experienced jurors reduces bias and thus improves judging quality in the short run. On the other, involving more inexperienced jurors in cases facilitates learning, thus nurturing a more sustainable juror pool in the long run. We further explore this trade-off in Section 6.

6. The Impact of In-group Bias on Case Outcomes

Thus far, we have established that in-group bias affects how crowd jurors cast their votes. However, as Taobao adopts a simple majority ruling system, and the final outcome of each case is determined by aggregating votes from a group of jurors, it is unclear how the impact of in-group bias on individual votes influences the final case outcomes. In this section, we quantify this impact using a simulation model that combines the empirical results in Section 5 and juror dynamics.

6.1. Simulation Setting

We construct a simulation process over a period of 500 days. The process captures juror and case heterogeneity, new juror joining and existing ones leaving, experience accumulation, as well as the juror voting behavior based on our empirical findings. We run 20 replications to ensure that the standard error is small. We provide an outline of the model as follows and present the technical details in Appendix F.

At the outset, we generate the initial juror pool with 35,308 enrolled jurors with characteristics (e.g., experience, buyer/seller status) based on the empirical distribution in our sample. At the beginning of each day, 190 new users join the juror pool, and a random number of existing jurors leave based on our observation that more experienced jurors are more likely to continue participation. Each juror is matched with a daily capacity (the maximum number of cases a juror participate in one day), which is also a function of the juror's experience per the data. For example, empirically, the least experienced jurors are approximately 70 times more likely to leave the platform than their most experienced peers, and their daily capacity is also less than 10% of the experienced ones.

At each day, the platform releases a number of cases for judging (1000 in our baseline scenario). We capture case heterogeneity by randomly assigning a case with the baseline probability that the seller will win according to the empirical distribution of the case fixed effects from our regression. As in practice, these cases are released in batches (50 cases in a batch as the baseline). For each case batch, we randomly sample a sub-set of jurors using experience-adjusted weights. We then generate a random response time for each sampled juror based on judging experience, capturing our observation that more experienced jurors respond faster. Jurors with remaining daily capacity vote in the order of their response time. The vote each juror cast is based on our regression result (Column 1 in Table 5), incorporating the impact of juror experience and buyer/seller status on voting preference.[21] By putting together the response time and the votes, the case is decided and closed based on the majority rule. All voting jurors earn 10 experience points by participating in one case. Based on the 7-vote-majority policy, this simulation process generates a juror dynamics that are similar to what we observe in practice. For example, the mean experience points on a vote-weighted basis is approximately 194,000 and the median vote is cast by someone with 131,000 experience points, both similar to what we observe in the sample (Table 1).

[21] To make the predicted probability more analogous to what we would get from a binary response model, we translate the OLS predicted probability to an estimated logistic distribution probability by applying the linear discriminant model correction (Allison et al. 2020).

6.2. Bias under the 7-Votes-Majority Policy

Based on the above simulation model, we first examine the impact of in-group bias on case outcomes under the 7-votes simple majority policy, which the Public Jury adopted during the majority of our sample period. To quantify this impact, we note that although our empirical findings have confirmed the existence of in-group bias, we cannot determine where the unbiased judgement lies except that it is likely to be in between the judgement of the seller and that of the buyer. To tackle this problem, we construct a conservative assessment of this impact by comparing three simulated scenarios.

[Mixed]: Treating jurors' buyer and seller status as it is, this scenario approximates the case outcomes obtained in practice when jurors include both sellers and buyers.

[All-B]: Treating all jurors as buyers, this scenario approximates the case outcomes when all in-group bias resides on the seller side (that is, buyer jurors yield unbiased results).

[All-S]: Treating all jurors as sellers, this scenario approximates the case outcomes when all in-group bias resides on the buyer side.

By comparing [Mixed] with [All-B], we obtain the estimated impact of in-group bias on case outcomes assuming in-group bias resides entirely on the seller side (*Bias-S*). Similarly, by comparing [Mixed] with [All-S], the estimated impact of in-group bias assumes that all in-group bias resides on the buyer side (*Bias-B*). In practice, it is more likely that both sellers and buyers are biased, and their impacts will hence (partially) cancel out. Thus, the two bias estimates above offer two conservative estimates of the actual magnitude of in-group bias in practice. Note that as we model the jurors' dynamic progress during the process, we only change each juror's buyer/seller status when simulating their vote on the current case, but then revert this status to their original state as they evolve over time.

The result based on the baseline scenario is presented in the first row of Table 6. As shown, the seller side wins approximately 38% of cases, similar to what we observe in the data. The estimated impact of (sole) seller bias is 0.309%. That is, if the in-group bias we observed empirically resides entirely on the seller side (i.e., buyers are un-biased), on average, 3 out of 1,000 cases are mistakenly decided in favor of the seller due to the existence of such in-group bias. Similarly, the estimated impact of (sole) buyer bias is 2.17%, which is greater than that of seller bias. This is intuitive: as the majority of jurors are buyers, if the in-group bias solely resides on the buyer side, it will have a larger impact on the case outcome. However, even under this conservative assumption, in-group bias will only affect only a small fraction of case outcomes. This is mainly due to two reasons. First, the vast majority of votes are cast by experienced jurors, who have a small or possibly negligible in-group bias. Second, although our empirical results have identified in-group bias across all cases, such in-group bias is most likely going to influence the outcome of those cases decided within a

Table 6 The Impact of In-group bias on Case Outcomes under the 7-Vote Majority Policy

Scenario	Outcome (%)	Bias-S (%)	Bias-B (%)	Notes
Baseline	37.983	0.309	2.166	Parameters summarized in Section 6.1.
	(0.096)	(0.004)	(0.012)	
Uniform case FEs	47.924	0.474	3.122	Case fixed effects drawn from a Uniform[0,1]
	(0.073)	(0.008)	(0.020)	distribution.
Random caseload	38.017	0.324	2.311	Daily caseload drawn from the empirical
	(0.088)	(0.004)	(0.023)	distribution in our sample.
Reduced batch size	37.957	0.241	1.783	The size of each case batch is reduced
	(0.048)	(0.004)	(0.017)	by half (from 50 to 25).
Inflated case points	37.973	0.230	1.732	A juror is awarded 25 points after completing
	(0.089)	(0.005)	(0.015)	on a case (instead of 10 in the baseline).

Notes. The standard error is reported in the parenthesis. *Outcome* is the fraction of cases ruled in favor of the seller under actual juror status ([Mixed]). *Bias-S* is the difference between the fraction of cases that the seller win under the [Mixed] scenario and [All-B]. *Bias-B* is the difference between the fraction of cases that the seller win under the [All-S] scenario and [Mixed].

thin margin. For those "clear-cut" cases, even if in-group bias could affect a vote or two, it will not be sufficient to alter the final outcome of the case. By including case fixed effects in the model, we simulate a mix of ambiguous and clear-cut cases according to the empirical distribution, limiting the number of cases whose outcome could be affected by in-group bias. This point is echoed by the second scenario that we examine, where we draw case fixed effects used in the simulation from a Uniform[0,1] distribution, which by assumption includes more ambiguous cases than the empirical distribution. As shown in Row 2, the bias affects the outcome of 3.12% of cases, an increase of 44% from the baseline. While this is still a relatively small amount, it does suggest that the impact of in-group bias could be more pronounced when more cases are ambiguous.

To assess the sensitivity of the estimated impact of in-group bias, we perturb a number of parameters in the simulation. First, we find that if we have a fluctuating daily case load as in the actual data (Row 3), in-group bias exhibits a modest 6% increase. This is because when the workload increases, more experienced jurors are likely to have exhausted their capacity and the average juror becomes more inexperienced. We also consider a smaller batch size, that is, the system will distribute smaller packs of cases with higher frequency (Row 4). Doing so further reduces the impact of in-group bias by 19% as experienced jurors, who will typically respond first, will vote on more cases. Finally, we note that our sample only covers one case category in the Public Jury system, and jurors could in practice accumulate experience points by judging other types of cases. This is partly accounted for through our initial point distribution, which views the point distribution as of the start of the sample. To further incorporate this feature in juror growth

dynamics, we increase the experience points gained by judging one case from 10 to 25 (Row 5). This captures the empirical observation that for the median juror-day, jurors gain 1.5 points through other categories for every point earned by judging transactional dispute cases. Not surprisingly, by allowing jurors to accumulate experience points faster, the impact of in-group bias further drops.

6.3. Policy Improvement

The above simulation confirms that under the baseline policy that Taobao adopted, in-group bias has limited impact on case outcome. Next, we examine if other voting policies and managerial interventions could help further reduce this impact, while at the same time better nurturing a more sustainable juror base. The results are presented in Table 7. In this table, we present the performance of these policies relative to the baseline 7-vote-majority policy.[22]

Table 7 Policy comparison (relative difference with respect to the 7-vote policy, in percentage)

	16-vote Majority	Dynamic (Experienced)	Dynamic (Mixed)	Inexperienced Jurors First	Increasing Enrollment
Panel A: Case Outcome					
Bias-S	−4.53	−70.6	−63.1	10.0	15.2
Bias-B	−17.4	−41.1	−39.4	54.4	17.4
Panel B: Voting Behavior					
Avg. # votes per case	128.2	−0.072	0.108	0.456	0.132
Avg. exp point by vote	−22.3	7.37	−0.157	−67.5	−5.91
Median exp point by vote	−32.5	18.5	−1.04	−82.8	−13.7
Panel C: Juror Experience at the end of the Simulation (Day 500)					
# Active Jurors	0.49	0.20	0.34	0.93	137.4
Avg. exp point	19.3	1.02	0.68	−3.18	−57.6
# Juror with Exp Level ≥ 3	8.43	1.01	1.59	5.31	1.17
# Juror with Exp Level ≥ 4	10.5	1.07	1.80	5.01	1.13

Notes. Relative difference is calculated as the difference between the quantity under the focal policy and that under the 7-votes majority policy normalized by the quantity under the 7-votes majority policy.

The first policy we consider is to increase the required majority from 7 to 16 (Column 1). This policy was briefly implemented on the Public Jury platform over our sample period. As shown, by more than doubling the number of jurors voting on a case (and thus resulting in a longer case resolution time), this policy only leads to a modest decrease in bias: 4.5% when assuming the bias solely resides on the seller side, and 17% on the buyer side. This is due to the drop in experience among participating jurors associated with the larger panel size requirement. Both the mean and

[22] We report the raw numbers and the corresponding standard errors in Appendix F (Table F.2).

median of the experience point by vote decline as we increase the panel size. This is consistent with the empirical finding illustrated in Figure 2 that juror experiences decline by vote ranking. This phenomenon reveals an important trade-off the platform faces when deciding panel size: while a larger panel size better reduces the effect of idiosyncratic decision-making of individual jurors (the "wisdom of the crowd" effect), its benefit can be largely offset by the decline of average judging quality. One additional benefit of a larger panel size is that it allows the jurors to accumulate experience faster. For example, the average experience points for a juror is approximately 20% higher under the larger panel policy, and the number of jurors with the Public Juror Experience Level 4 or above (with experience point 12,000 or above) increases by 10%.

Next, motivated by the relationship between experience and in-group bias, we design three policies that dynamically allocate cases among jurors. In the "Dynamic (Experienced)" policy (Column 2), we use the first five votes to decide whether the case is an ambiguous one (defined as the first five votes having a 2-3 split, accounting for approximately 18.3% of cases). If the case is determined as ambiguous, we exhibit this case only to experienced jurors (jurors with experience points more than 54,600, which corresponds to the threshold for the Public Juror Experience Level 5) subsequently.[23] By targeting these ambiguous cases, this policy significantly reduces the impact of in-group bias (70% in Bias-S, and 40% in Bias-B).

Despite its benefit in bias reduction, the above policy allocates more cases to experienced jurors, thus limiting the opportunity for inexperienced jurors to participate and grow. To better balance bias reduction and juror growth, the "Dynamic (Mixed)" policy (Column 3) makes the ambiguous cases (2-3 split among the first five votes) only available to experienced jurors, as in the previous policy, but at the same time allocate cases identified as "clear-cut" (defined as the first five votes having a 0-5 split, 51.8% cases are classified as "clear-cut") only to inexperienced jurors. Finally, cases in between (with 1-4 split among the first five votes) are made available to all jurors. This policy is almost equally effective at bias reduction as the "Dynamic (Experienced)" policy, but is better at nurturing inexperienced jurors. For example, the increase of the number of jurors with Experience Level 4 or above relative to the static 7-vote baseline is 1% under the "Dynamic (experienced)" policy, but 1.8% under the "Dynamic (Mixed)" policy.

To further explore the implications of further targeting at inexperienced jurors, we consider a policy we call "Inexperienced Jurors First" (Column 4). Under this policy, we restrict the first five jurors to be inexperienced, and only involves experienced jurors when the case is deemed ambiguous by the first five votes. As shown, while this policy helps more inexperienced jurors to become experienced, it suffers from a higher in-group bias (54% more than under the baseline).

[23] We have explored other experienced point cut-offs and we obtain similar results.

Finally, in addition to voting policies, we consider another management intervention: increasing new juror enrollment (e.g., through aggressive recruiting efforts). In this hypothetical scenario, we assume that the platform could double the number of users enrolled as new jurors every day (from 190 to 380). Interestingly, this policy actually aggravates the impact of in-group bias. The reason is that new jurors tend to be very active during a short period immediately after their enrollment into the Public Jury.[24] These new (and thus inexperienced jurors) could then crowd out more experienced jurors, thus lowering overall judging quality. Further, due to their high attrition rate, these new jurors do not translate to a larger pool of (relatively) more experienced jurors. In fact, the number of jurors with Experience Level 3 (or 4) and above under this policy is even lower than the Dynamic (Mixed) policy. These findings caution the platform operators of the possible negative impact of aggressively recruiting new jurors.

7. Concluding Remarks

As the first empirical study on crowd-sourced dispute resolution, our paper has several limitations. First, while our empirical analysis strongly suggests the existence of in-group bias, we cannot directly observe whether jurors exhibit any selection behavior. Should related data become available, we could obtain cleaner identification. Data on case characteristics, as well as evidence presented to the jurors, may also allow us to better understand the mechanism behind in-group bias and the moderators (e.g., experience). Further, our dataset is based on qualified Taobao users who volunteer to act as jurors, thus our findings of in-group bias may not apply to all users on the platform, or the general population.

This paper can also be extended along several different directions. For example, comparing the crowd's decision and the decision of experts (e.g., customer service representatives) could help us better understand the quality of crowd-judging beyond in-group bias.[25] It would also be interesting to quantify juror behavior on different types of platforms or in a different cultural context. Finally, a better understanding of how platform users perceive the legitimacy of different operational designs of crowd-judging could be a promising research direction.

Acknowledgments

The authors Professor Vishal Gaur, the Associate Editor and three referees for their construction comments. We thank Alibaba for providing crowd-judging data and institutional details. For their valuable input,

[24] For example, a new juror has a more than 75% probability to vote a case on the day that they are enrolled in the Public Jury. See Appendix F for more details.

[25] In the absence of any objectively "correct" decision for each case, we have conducted empirical analysis on vote consistency (whether a juror's vote on a case is consistent with the majority ruling). We find that as jurors gain experience, their vote consistency also increases, and part of this improvement is due to jurors converging with their in-group peers, suggesting that experience improves judging quality beyond reducing bias. See Appendix E for details.

the authors thank John Birge, Ruomeng Cui, Nitish Jain, Gillian Ku, Kamalini Ramdas, Nicos Savva, the participants of European TOM Seminar Series, the 2019 American Law and Economics Association Annual Conference, the HKU-Lingnan-Florida Platform Competition Conference, 14th Annual Conference on Empirical Legal Studies, and 23rd Annual Conference of the Society for Institutional and Organizational Economics, as well as workshop participants at University of Michigan, University of Southern California, George Mason University and Academia Sinicia.

References

Aksin, O Z., S. Deo, J. O. Jónasson, K. Ramdas. 2021. Learning from many: Partner exposure and team familiarity in fluid teams. *Management Science* **67**(2) 854–874.

Alibaba Inc. 2016. Alibaba ecosystem internet volunteers research report. Report on file with the authors.

Allison, P. D, R. A Williams, P. von Hippel. 2020. Better predicted probabilities from linear probability models. *Statistical Horizons* .

Antweiler, W., M. Z Frank. 2004. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance* **59**(3) 1259–1294.

Anwar, S., P. Bayer, R. Hjalmarsson. 2012. The impact of jury race in criminal trials. *The Quarterly Journal of Economics* **127**(2) 1017–1055.

Bakos, Y., C. Dellarocas. 2011. Cooperation without enforcement? A comparative analysis of litigation and online reputation as quality assurance mechanisms. *Management Science* **57**(11) 1944–1962.

Benjaafar, S., G. Kong, X. Li, C. Courcoubetis. 2019. Peer-to-peer product sharing: Implications for ownership, usage, and social welfare in the sharing economy. *Management Science* **65**(2) 477–493.

Bernstein, L. 1992. Opting out of the legal system: Extralegal contractual relations in the diamond industry. *The Journal of Legal Studies* **21**(1) 115–157.

Bimpikis, K., O. Candogan, D. Saban. 2019. Spatial pricing in ride-sharing networks. *Operations Research* **67**(3) 744–769.

Boh, W., S. A Slaughter, J A. Espinosa. 2007. Learning from experience in software development: A multilevel analysis. *Management science* **53**(8) 1315–1331.

Boudreau, K. 2010. Open platform strategies and innovation: Granting access vs. devolving control. *Management science* **56**(10) 1849–1872.

Boudreau, K. J, N. Lacetera, K. R Lakhani. 2011. Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management science* **57**(5) 843–863.

Budescu, D. V, E. Chen. 2014. Identifying expertise to extract the wisdom of crowds. *Management Science* **61**(2) 267–280.

Cachon, G. P, K. M Daniels, R. Lobel. 2017. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management* **19**(3) 368–384.

Cai, H., Y. Chen, H. Fang. 2009. Observational learning: Evidence from a randomized natural field experiment. *American Economic Review* **99**(3) 864–82.

Carter, E. R, M. C Murphy. 2015. Group-based differences in perceptions of racism: What counts, to whom, and why? *Social and Personality Psychology Compass* **9**(6) 269–280.

Chan, J., J. Wang. 2018. Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Science* **64**(7) 2973–2994.

Chen, D. L, T. J Moskowitz, K. Shue. 2016. Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics* **131**(3) 1181–1242.

Chen, Y., S. X. Li. 2009. Group identity and social preferences. *American Economic Review* **99**(1) 431–57.

Chod, J., N. Trichakis, S A. Yang. 2022. Platform tokenization: Financing, governance, and moral hazard. *Management Science* **68**(9) 6411–6433.

Cohen, M., K. Jiao, J. Serpa. 2020. The impact of IPOs on peer-to-peer lending platforms. *McGill University Working Paper* .

Cui, R., J. Li, D. J Zhang. 2020. Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. *Management Science* **66**(3) 1071–1094.

Cui, R., D. J Zhang, A. Bassamboo. 2019. Learning from inventory availability information: Evidence from field experiments on Amazon. *Management Science* **65**(3) 1216–1235.

Da, Z., X. Huang. 2020. Harnessing the wisdom of crowds. *Management Science* **66**(5) 1847–1867.

Del Duca, L. F, C. Rule, K. Rimpfel. 2014. eBay's De Facto Low Value High Volume Resolution Process: Lessons and Best Practices for ODR Systems Designers. *Arbitration Law Review* **6**(1) 204–219.

Dovidio, J. F, A. Love, F. MH Schellhaas, M. Hewstone. 2017. Reducing intergroup bias through intergroup contact: Twenty years of progress and future directions. *Group Processes & Intergroup Relations* **20**(5) 606–620.

Edelman, B., M. Luca, D. Svirsky. 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics* **9**(2) 1–22.

Egan, M. L, G. Matvos, A. Seru. 2018. Arbitration with uninformed consumers. Tech. rep., National Bureau of Economic Research.

Epstein, L., W. M Landes, R. A Posner. 2013. *The behavior of federal judges: a theoretical and empirical study of rational choice*. Harvard University Press.

Fan, G., X. Wang, H. Zhu. 2018. NERI index of marketization of China's provinces 2018 report. `https://www.pishu.com.cn/skwx_ps/bookdetail?SiteID=14&ID=10744111` (in Chinese, accessed on 11 November 2020).

Foerderer, J. 2020. Interfirm Exchange and Innovation in Platform Ecosystems: Evidence from Apple's Worldwide Developers Conference. *Management Science* **66**(10) 4772–4787.

Franke, N., E. Von Hippel. 2003. Satisfying heterogeneous user needs via innovation toolkits: the case of Apache security software. *Research policy* **32**(7) 1199–1215.

Gawronski, B., G. V Bodenhausen. 2006. Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological bulletin* **132**(5) 692.

Greene, W. 2004. Fixed effects and bias due to the incidental parameters problem in the Tobit model. *Econometric reviews* **23**(2) 125–147.

Greenstein, S., F. Zhu. 2012. Is Wikipedia biased? *American Economic Review* **102**(3) 343–48.

Greenstein, S., F. Zhu. 2018. Do experts or crowd-based models produce more bias? Evidence from Encyclopedia Britannica and Wikipedia. *MIS Quarterly* **42**(3) 945–959.

Greenwald, A. G, M. R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review* **102**(1) 4–27.

Greif, A. 1993. Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *American Economic Review* 525–548.

Gurvich, I., M. Lariviere, A. Moreno. 2019. Operations in the on-demand economy: Staffing services with self-scheduling capacity. *Sharing Economy*. Springer, 249–278.

Hewstone, M., M. Rubin, H. Willis. 2002. Intergroup bias. *Annual review of psychology* **53**(1) 575–604.

Huang, Y., P. Vir Singh, K. Srinivasan. 2014. Crowdsourcing new product ideas under consumer learning. *Management science* **60**(9) 2138–2159.

Huckman, R. S, G. P Pisano. 2006. The firm specificity of individual performance: Evidence from cardiac surgery. *Management Science* **52**(4) 473–488.

Jordan, J. 2015. Managing tensions in online marketplaces. URL `https://techcrunch.com/2015/02/23/managing-tensions-in-online-marketplaces/`.

Klerman, D M. 2022. Bias in choice of law: New empirical and experimental evidence. *Journal of Institutional and Theoretical Economics* (22-27). Forthcoming.

Kremer, I., Y. Mansour, M. Perry. 2014. Implementing the "Wisdom of the Crowd". *Journal of Political Economy* **122**(5) 988–1012.

Larrick, R. P, J. B Soll. 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science* **52**(1) 111–127.

Lee, W. K., Y. Cui. 2022. Should Gig platforms decentralize dispute resolution? *Working Paper, Cornell University, Available at SSRN 3719630* .

Levitt, S. D, J. A List, C. Syverson. 2013. Toward an understanding of learning by doing: Evidence from an automobile assembly plant. *Journal of Political Economy* **121**(4) 643–681.

Mejia, J., C. Parker. 2021. When transparency fails: Bias and financial incentives in ridesharing platforms. *Management Science* **67**(1) 166–184.

Mollick, E., R. Nanda. 2015. Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Management Science* **62**(6) 1533–1553.

Musalem, A., M. Olivares, D. Yung. 2019. Balancing agent retention and waiting time in service platforms. *Available at SSRN (3502469)* .

Papanastasiou, Y., S A. Yang, A. H. Zhang. 2022. Improving dispute resolution in two-sided platforms: The case of review blackmail. *Management Science* Forthcoming.

Parker, G., M. Van Alstyne. 2018. Innovation, openness, and platform control. *Management Science* **64**(7) 3015–3032.

Posner, R. A. 2010. *How judges think*. Harvard University Press.

Price, J., J. Wolfers. 2010. Racial discrimination among NBA referees. *The Quarterly Journal of Economics* **125**(4) 1859–1887.

Quillian, L. 1995. Prejudice as a response to perceived group threat: Population composition and anti-immigrant and racial prejudice in Europe. *American Sociological Review* 586–611.

Ramdas, K., K. Saleh, S. Stern, H. Liu. 2018. Variety and experience: Learning and forgetting in the use of surgical devices. *Management Science* **64**(6) 2590–2608.

Rehavi, M M., S. B Starr. 2014. Racial disparity in federal criminal sentences. *Journal of Political Economy* **122**(6) 1320–1354.

Rule, C., C. Nagarajan. 2010. Leveraging the wisdom of the crowds: the eBay community court and the future of online dispute resolution. *ACResolution 2 (2)* **4–7**.

Sauermann, H., C. Franzoni. 2015. Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences* **112**(3) 679–684.

Shafer, S. M, D. A Nembhard, M. V Uzumeri. 2001. The effects of worker learning, forgetting, and heterogeneity on assembly line productivity. *Management Science* **47**(12) 1639–1653.

Shayo, M., A. Zussman. 2011. Judicial ingroup bias in the shadow of terrorism. *The Quarterly Journal of Economics* **126**(3) 1447–1484.

Sherif, M., O.J. Harvey, B.J. White, W.R. Hood, C.W. Sherif. 1961. *Intergroup conflict and cooperation: The Robbers Cave experiment*. University of Oklahoma Press.

Sommers, S. R, M. I Norton. 2006. Lay theories about White racists: What constitutes racism (and what doesn't). *Group Processes & Intergroup Relations* **9**(1) 117–138.

Tajfel, H., M. G Billig, R. P Bundy, C. Flament. 1971. Social categorization and intergroup behaviour. *European journal of social psychology* **1**(2) 149–178.

Tajfel, H., J. C Turner, W. G Austin, S. Worchel. 1979. An integrative theory of intergroup conflict. M. J. Hatch, M. Schultz, eds., *Organizational identity: A reader*. Oxford University Press, 56–65.

Terwiesch, C., Y. Xu. 2008. Innovation contests, open innovation, and multiagent problem solving. *Management science* **54**(9) 1529–1543.

Tversky, A., D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* **185**(4157) 1124–1131.

Van Loo, R. 2016. The corporation as courthouse. *Yale J. on Reg.* **33** 547.

Xu, Y., M. Armony, A. Ghose. 2021. The interplay between online reviews and physician demand: An empirical investigation. *Management Science* **67**(12) 7344–7361.

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 juror's 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

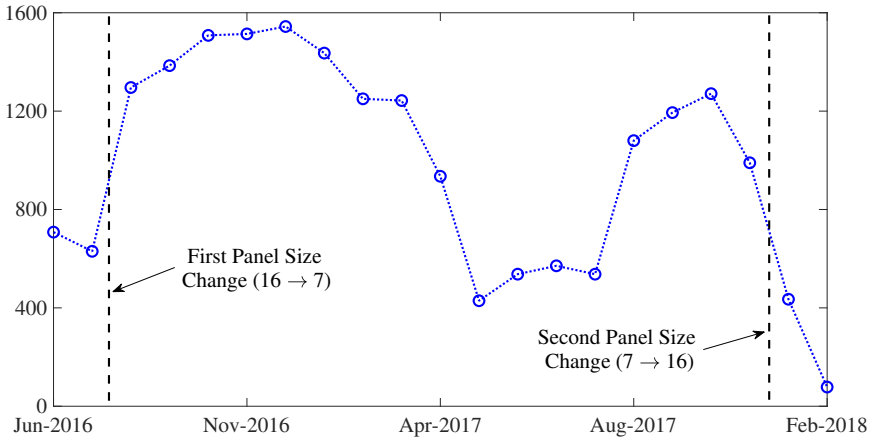
Experience Level	Survival by Month (%)	Daily Attrition Probability (%)
1	51.46	2.190
2	62.43	1.558
3	74.73	0.966
4	86.00	0.502
5	92.45	0.261
6	95.82	0.142
7	98.46	0.052
8	99.05	0.032

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,

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

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 E.2 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.