cbess
Centre for Behavioural and Experimental Social Science

University of East Anglia

2023-01

Strategic Thinking in Jury Decisions: An Experimental Study

Can Çelebi
Stefan P. Penczynski

www.uea.ac.uk/cbess

Strategic thinking in jury decisions: an experimental study*

Can Çelebi†
Stefan P. Penczynski‡

December 8, 2023

Theoretical work by Feddersen and Pesendorfer (1998) has shown how strategic voting undermines the intuition that unanimous voting eliminates convictions of innocent defendants. We set up a level-k model of jury voting and experimentally investigate strategic thinking with an experimental design that uses intra-team communication. Looking at juries using the unanimity rule, we show that the jury performance depends on the strategic sophistication of jury members, which in turn depends on the complexity of the task at hand.

Keywords: Jury voting, levels of reasoning, strategic voting.

JEL Classification: D72, D83.

*We thank Felix Gundert and Stefan Schaefauer for excellent research assistance. The paper has benefitted from comments received from the audience at the 2018 UEA Behavioural Game Theory Workshop and Amir Jafarzadeh, Andrea Marietta Leina, and Janithe Siriwardana.

†Department of Economics, University of Mannheim. Max Planck Institute for Intelligent Systems in Tübingen, cnelebi@gmail.com.

‡School of Economics and Centre for Behavioural and Experimental Social Science (CBESS), University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom, s.penczynski@uea.ac.uk, Tel. +44 1603 59 1796.

1. Introduction

Predicated on the assumption that jurors vote non-strategically, the unanimity rule in juries has traditionally been seen as a safeguard against convicting the innocent. Building upon the work of Austen-Smith and Banks (1996), Feddersen and Pesendorfer (1998, FP) challenged this perspective and showed that, under the a priori assumption that a defendant is equally likely to be innocent or guilty, an equilibrium with the unanimity rule can encompass strategic voting. In contrast to informative voting, strategic voting might not simply reveal private information.

FP proved that the presence of strategic voting leads to higher error rates in convicting the innocent and acquitting the guilty compared to juries under simple majority or supermajority rules, where strategic voting is absent in equilibrium under the same initial assumptions. Furthermore, they highlighted that the issue of an increased rate of convicting the innocent under unanimity voting becomes notably exacerbated with larger jury sizes.

Guarnaschelli et al. (2000, GMP) experimentally confirmed the presence of strategic voting under unanimity voting, but also documented results that to some extent contradict FP's claims on error rates. They sought to reconcile this partial discrepancy between theory and data by utilizing the Quantal Response Equilibrium (QRE) model's statistical nature (McKelvey and Palfrey, 1995).

We introduce an alternative non-equilibrium framework, using level-k modeling, to address this discrepancy and offer a fresh perspective on jury voting under the unanimity rule, both theoretically and experimentally. Despite a large literature on strategic thinking and heterogeneous level-k reasoning that includes work on auctions, social learning and other related settings (Crawford et al., 2013), to the best of our knowledge the precise cognitive processes of strategic voting and their concrete influence on jury accuracy have not yet been studied.

We model the jury voting via iterative best responses. We begin with the assumption that all level-0 players vote uninformatively. A level-1 player best-responds to level-0 players by voting her signal (informative voting); and a level-2 player best-responds to level-1 players by voting "guilty" irrespective of her signal (strategic voting). The intuition is that informative voting with imprecise signals will otherwise never lead to convictions under the unanimity rule. With strategic voting being inherently uninformative, level-3 players revert to informative voting in response. This establishes a cyclical best-response pattern: higher odd-level players respond to strategic voting with informative voting and higher even-level players counteract informative voting

with strategic voting.

We attained nuanced insights into individual reasoning beyond mere vote observation by employing an intra-team communication protocol that provides written accounts of subjects' decision justifications (Burchardi and Penczynski, 2014; Penczynski, 2017). These accounts substantiated the existence of the modelled level-k reasoning types within our sample and allowed us to differentiate among overlapping voting behaviors of various level-k types, that arises due to our model's cyclical best-response characteristic. Additionally, this approach enhanced the robustness of identifying strategic voters by screening out level-0 players who might otherwise be misidentified as strategic voters if categorizations were exclusively based on observed votes.

In a 2×2 within-subject experimental design, we investigate strategic voting across two jury sizes ($n = 3, 6$) and two sampling methods for the information signal (with and without replacement). While our theoretical model predicts that specific level-k types do not change their behaviour when jury sizes or signal sampling methods vary, both variations are predicted to affect the Nash equilibrium (NE) behavior and are expected to influence the complexity of the strategic task.

Given a specific level-classification, the observed voting behavior closely mirrors the theorized behavior and remains largely unaffected by treatments. Yet, these treatments influence the aggregate voting behavior by altering the distribution of strategic sophistication. This influence partly explains the discrepancies between observed jury accuracy and the optimal accuracy predicted by NE. Crucially, the level distributions are primarily affected by the strategic complexity introduced by the treatment variation, so that the aggregate voting behavior not necessarily aligns with the expected changes under optimal voting.

We find results in line with GMP, who found no significant difference in the frequency of strategic voting between jury sizes of 3 and 6. Both studies rather find a significant decrease in convicting an innocent, and an increase in acquitting a guilty, contrary to the NE predictions. We can explain this with the informative votes of less sophisticated level-1 types, which become more frequent in the more complex setting with $n = 6$ jurors. We similarly find that the sampling without replacement seems less complex and leads to higher strategic sophistication compared to sampling with replacement.

In light of these results, we propose that a model of heterogeneous types of voters such as the level-k model is better suited to understand and predict jury performances across a number of different settings than NE or QRE because it represents the underlying cognitive processes of strategic voting.

Our results relate the setting of jury voting to other settings, in which the plurality of types has been identified as important for a good overall outcome. For example, in social learning, the heterogeneity of types improves upon the inefficiencies of fully rational behaviour – information cascades and herding – thanks to occasional and private-information-revealing level-1 decisions (Penczynski, 2017). Furthermore, the fact that the observed level distribution still leads to a good jury performance resembles the coordination "magic" in market-entry games, in which heterogeneous beliefs seem to provide a useful mechanism of sorting otherwise homogeneous players into market entrants and non-entrants (Rapoport et al., 1998; Camerer et al., 2004).

Due to the observed influence of the treatment variations on the level-k distribution, our work also relates to the idea of endogenous depth of reasoning (Agranov et al., 2012; Alaoui and Penta, 2016). If the strategic sophistication is the result of a cost-benefit analysis of additional steps of reasoning as modeled by Alaoui and Penta (2016), our analysis suggests that a larger number of possible signal realizations within the jury, be it due to a larger jury or to sampling without replacement, increases the complexity of the game and the cost of deliberation and hence lowers the average observed sophistication.

Our results suggest that conditioning on being pivotal – casting a decisive vote that determines the jury decision – in voting is much less inhibitive of strategic sophistication than conditioning on bidding highest in first-price auctions (Eyster and Rabin, 2005; Crawford and Iriberri, 2007; Li, 2017). In auctions, an intra-team communication analysis suggests that "subjects may actually have problems to form even a basic belief" as only 15% of subjects deliberate the other players' decision and thus qualify for anything other than level-0 (Koch and Penczynski, 2018, p. 79). In contrast, the level distribution in voting is very similar to commonly found distributions in other settings.

Incorporating sampling without replacement in our experiment is inspired by Rabin (2002), who models inference via a sampling process without replacement in order to reflect the common belief in the law of small numbers (Tversky and Kahneman, 1974). Similar to our motivation, the increased difficulty of dealing with independent signal and the simplification of draws that are more "representative" of the urn composition have led Grimm and Mengel (2020) to use sampling without replacement in their experiment. Our results support the intuition that this sampling is easier to understand.

2. Theory

2.1. Model Setup

Consider a game with n jurors. Nature determines the state of the world to be red or blue, $S \in \{R, B\}$, where each state is equally likely to occur. The realization of S is not observable by the jury members. After S is determined, each juror receives a private red or blue signal, $s \in \{r, b\}$. Signals are informative, as the colors for state and signal coincide with $p \in (\frac{1}{2}, 1)$, and differ with probability $1 - p$.

Assume the state of the world is represented by the color of an urn and the jury members' signals are balls drawn from the urn. Given $p > \frac{1}{2}$, relatively more blue (red) balls are in the blue (red) urn.

We distinguish between two types of sampling of the balls: without replacement (O) and with replacement (W). In O, the private signal observed by the juror is dependent on the private signals observed by the other $n-1$ jury members. More specifically, p is a realization of a hypergeometric random variable. In contrast, in W, the private signal observed by the player is independent of the private signals observed by the other jury members, and p is a realization of a Bernoulli random variable.

After each juror receives her private signal, she votes as a part of the jury to correctly guess the true state of the world with vote $v \in \{R, B\}$. The votes of the n jurors are aggregated into a jury decision $\hat{v} \in \{R, B\}$ according to the unanimity voting rule: the jury decides for the red urn if and only if all the jury members vote red, and decides for the blue urn otherwise.

Given this notation, the probabilities of convicting an innocent defendant and acquitting a guilty defendant are respectively represented as $Pr(\hat{v} = R|S = B)$ and $Pr(\hat{v} = B|S = R)$. For the ease of notation, in the rest of the paper, we will refer to these error rates as $Pr(R|B)$ and $Pr(B|R)$.

Every juror has the same payoff function $U(\hat{v}, S)$ and is assumed to receive $\pi = 0$ if the jury correctly guesses the color of the urn; and bears a cost of $q \in (0, 1)$ for wrongly identifying a blue urn as red, and a cost of $(1 - q)$ for wrongly identifying a red urn as blue. In summary, we have:

$$U(R, R) = U(B, B) = \pi = 0$$
$$U(R, B) = -q$$
$$U(B, R) = -(1 - q)$$

Given these payoffs, the parameter q defines a juror's boundary for reasonable doubt. A juror who believes the defendant to be guilty with probability higher than q will

strictly prefer to convict the defendant. A greater value of q indicates that the juror is more tolerant of the risk of acquitting a guilty defendant compared to the potential harm of convicting an innocent.

2.2. Nash Equilibrium

Define $\sigma(s)$ as the probability to vote red given signal s. FP show the existence of a mixed strategy equilibrium in which every juror votes red with some positive probability, $\sigma(b)$, when her signal is blue; and always vote red, $\sigma(r) = 1$, when her signal is red. We adapt their relevant findings – for sampling with replacement – to our terminology and summarize them in the following proposition:

Proposition 1
Given the signals are independent from each other, the unique symmetric mixed strategy equilibrium is defined as

$$\sigma(r) = 1, \tag{1}$$

$$\sigma(b) = \frac{Kp - (1-p)}{p - K(1-p)}, \tag{2}$$

where

$$K = \left(\frac{(1-q)(1-p)}{qp} \right)^{\frac{1}{n-1}}. \tag{3}$$

Moreover, the probability of an incorrect jury decision to vote red when the true state is blue, $Pr(R|B)$, and the probability to vote blue when the true state is red, $Pr(B|R)$, are defined as

$$Pr(R|B) = (\rho_B)^n, \tag{4}$$

$$Pr(B|R) = 1 - (\rho_R)^n, \tag{5}$$

where

$$\rho_R = p\sigma(r) + (1-p)\sigma(b) \text{ and} \tag{6}$$

$$\rho_B = (1-p)\sigma(r) + p\sigma(b) \tag{7}$$

are the probabilities that a juror votes red for the respective states of the world, R and B.

Proof. See Feddersen and Pesendorfer (1998).[1] ■

Proposition 1 defines an equilibrium solution for the case where the private signals are drawn independently (W). We provide the mixed strategy equilibrium for the hypergeometric case (O) in the following corollary.

Corollary 1.1 *Given the signals are drawn from the hypergeometric distribution, the unique symmetric mixed strategy equilibrium is defined as*

$$\sigma(r) = 1, \tag{8}$$

$$\sigma(b) = \left(\frac{qp}{(1-p)(1-q)}\right)^{\frac{1}{n(1-2p)}}. \tag{9}$$

Moreover, the probability of an incorrect jury decision to vote red when the true state is blue, $Pr(R|B)$, and the probability to vote blue when the true state is red, $Pr(B|R)$, are defined as:

$$Pr(R|B) = (\rho_B)^n, \tag{10}$$

$$Pr(B|R) = 1 - (\rho_R)^n \tag{11}$$

where

$$\rho_R = \sigma(r)^{pn}\sigma(b)^{(1-p)n} \text{ and} \tag{12}$$

$$\rho_B = \sigma(r)^{(1-p)n}\sigma(b)^{pn}. \tag{13}$$

Proof. See Appendix A.1. ■

2.3. Best responses

Define α^s as the juror's belief about her probability of being pivotal given the signal s and conditional on state R, and define β^s for this belief conditional on state B. In the following proposition, we identify the conditions for informative and strategic voting under any beliefs α^s and β^s.

Proposition 2

For every juror i, define $u_i(\cdot)$ as the utility given her vote. Then, we have:

$$\mathbb{E}(u_i(\sigma_i(r)=1)) = ((1-q)\alpha^r p - q\beta^r(1-p)) + \mathbb{E}(u_i(\sigma_i(r)=0)) \tag{14}$$

$$\mathbb{E}(u_i(\sigma_i(b)=0)) = (q\beta^b p - (1-q)\alpha^b(1-p)) + \mathbb{E}(u_i(\sigma_i(b)=1)) \tag{15}$$

[1] More specifically, see pages 24-26 and Appendix A in FP for the proposition and its proof. For a brief overview of the explicit functional forms see pages 408-409 in GMP. For a brief discussion of the strategic voting and pivotality see pages 376-377 and 386-387 in Coughlan (2000) and pages 35-39 in Austen-Smith and Banks (1996).

Proof. See Appendix A.2. ■

Using Proposition 2, we identify the conditions under which a juror votes informatively or strategically in the following corollary.

Corollary 2.1
Assume $\alpha^s + \beta^s > 0$ *for* $s \in \{r, b\}$*, and define* $w = \frac{1-q}{q}$ *then a juror votes her signal (informative voting) if and only if* $p > \frac{w\alpha^b}{w\alpha^b+\beta^b}$ *and* $p > \frac{\beta^r}{w\alpha^r+\beta^r}$*; and a juror always votes red (strategic voting) if and only if* $\frac{w\alpha^b}{w\alpha^b+\beta^b} > p > \frac{\beta^r}{w\alpha^r+\beta^r}$.

Proof. See Appendix A.3. ■

The assumption $\alpha^s + \beta^s > 0$ assures that a juror is pivotal in at least one state of the world.

Notice that if we assume that either type of errors are equally costly as in our experimental setup, then we have $w = 1$; and the boundary conditions for informative and strategic voting become solely dependent on the pivotality probabilities. In the continuation of the paper, for the ease of notation and to be in line with our experimental setup, we will assume that $U(R, B) = U(B, R)$ and hence set q to $\frac{1}{2}$.

Moreover notice that if we further assume that $\sigma(r) \geq \sigma(b)$, then since $p > \frac{1}{2}$, it is trivial to show $\alpha^s \geq \beta^s$ for $s \in \{r, b\}$. Given $\alpha^s \geq \beta^s$ for $s \in \{r, b\}$, the inequality $\frac{\beta^r}{\alpha^r+\beta^r} \leq \frac{1}{2} < p$ is always satisfied, and the inequality $\frac{\beta^r}{\alpha^r+\beta^r} > p > \frac{1}{2}$ is never satisfied. As a result, informative and strategic voting conditions respectively simplifies to $\frac{\alpha^b}{\alpha^b+\beta^b} > p$ and $\frac{\alpha^b}{\alpha^b+\beta^b} < p$.

2.4. Level-k Modeling

Consider a model of heterogeneous types of strategic reasoning, with types $k \in \mathbb{N}^0$, who apply k iterated best responses to a level-0 belief (See Nagel, 1995; Stahl and Wilson, 1995; Crawford et al., 2013). The types are distributed in the population according to the level-k distribution $d(k)$, and each level-k juror believes all other jury members to be level-$(k-1)$. Furthermore, each juror chooses strategy $\sigma_k(s)$ and believes to be pivotal with probabilities α_k^s and β_k^s for states R and B respectively.

We assume a level-0 juror to vote uninformatively and hence independently of her signal, $\sigma_0(r) = \sigma_0(b) > 0$.[2] With $\sigma_0(s) > 0$ for $s \in \{r, b\}$, we avoid a trivial setup and assure the level-1 juror to be pivotal with positive probability.

[2]We can relax this assumption by introducing some degree of informativeness to level-0 voting, and maintain the same level-1 predictions (see Appendix A.6 for details). In addition, notice that $\sigma_0(r) = \sigma_0(b) > 0$ entails the case for level-0 jurors to vote randomly, i.e. $\sigma_0(s) = \frac{1}{2}$ for $s \in \{r, b\}$.

Given the uninformative level-0 voting, a level-1 juror's pivotality is independent of the state. Hence, a level-1 juror's expected utility solely depends on her signal's informativeness. Consequently, given the received signal has some degree of informativeness (i.e. $p > \frac{1}{2}$), a level-1 juror votes the same color as her signal.[3]

A juror's expected utility for either choice of vote depends both on her received signal's strength, p, and her perceived pivotality for a given state, α^s or β^s. Given the asymmetric nature of unanimous voting, a level-2 juror believes to be less pivotal if the true state is B. Hence if her perceived pivotality under state R is relatively high enough to offset her received blue signal's strength ($\alpha^b(1-p) > p\beta^b$), she will vote red upon a blue signal.[4]

Because a level-2 juror votes strategically, i.e. always votes red, her action becomes as uninformative as a level-0 juror. Hence, a level-3 juror best responds in the same way a level-1 juror does and votes informatively.

Proposition 3

Consider the case where each juror receives an independent private signal. $\forall n > 2$*, jury members at levels* $\{1, 3, ...\}$ *vote informatively; and jury members at levels* $\{2, 4, ...\}$ *vote strategically.*

Proof. See Appendix A.4. ■

For the case where the private signals are dependent (O), given $n \geq \frac{2}{1-p}$,[5] a level-2 juror never believes to be pivotal as she believes that there is always a level-1 juror in the group that receives the blue signal and votes blue. Therefore, for $n \geq \frac{2}{1-p}$, a level-2 juror is indifferent between voting blue and red.

To eliminate this indifference, we introduce a small error term, ϵ, into the juror's belief about other players' choices. Specifically, we assume that every juror believes with some small probability $\epsilon > 0$ that every other juror votes the color other than the

[3] If we relax the assumption that the jury errors are equally costly and assume without loss of generality that convicting the innocent is more costly for the juror than acquitting the guilty ($q > \frac{1}{2}$), then the degree of the received signal's informativeness needs to be higher than the juror's threshold of reasonable doubt, $p > q$, in order for her to vote informatively.

[4] If we relax the assumption that the jury errors are equally costly, then strategic voting will also be dependent on the relative costs of acquitting a guilty and convicting an innocent. For instance, if the jury members are primarily concerned with not convicting an innocent, then the threshold of reasonable doubt can be high enough to offset the relatively small probability of being pivotal under the blue state and a "level-2" juror will not vote strategically. On the other hand, given the juror is not pivotal in the blue state, $\beta^b = 0$, (as can be the case under without replacement sampling), then even if she has a very high threshold of reasonable doubt, she will still prefer to vote strategically.

[5] Given $p = \frac{2}{3}$, for $n \geq 6$.

color they are expected to vote, i.e. makes a mistake.[6] The introduced noise enables us to parallel the predictions we made in Proposition 3, for O in the following proposition.

Proposition 4

Consider the case where each juror receives a hypergeometrically dependent private signal. Assume that each juror believes other jury members to make a mistake with some probability ϵ such that:

$$\epsilon < \frac{1}{1 + \left(\frac{p}{1-p}\right)^{\frac{1}{(2p-1)n}}}$$

Then $\forall n > 2$, jury members at levels $\{1, 3, ...\}$ vote informatively, and jury members at levels $\{2, 4, ...\}$ vote strategically.

Proof. See Appendix A.5. ■

On the basis of Propositions 3 and 4, Table 1 spells out the level-k predictions for level-0 to level-3 by sampling method, jury size and signal. Notably, each level-k type behaves the same across these different aspects. Differences in $\sigma(s)$ across treatments can therefore only result from changes in the level-k distribution $d(k)$.

	O				W			
	$n=3$		$n=6$		$n=3$		$n=6$	
$d(k)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$
$k=0$	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
$k=1$	0	1	0	1	0	1	0	1
$k=2$	1	1	1	1	1	1	1	1
$k=3$	0	1	0	1	0	1	0	1

Table 1: Level-k predictions for $\sigma_k(s)$.

Given the importance of $d(k)$, asking for an ideal level-k distribution becomes insightful. Specifically, which ideal distribution $d^*(k)$ minimizes the aggregate probability of a jury's errors, $Pr(R|B) + Pr(B|R)$? We answer this question in an optimization problem that features three distinct types, level-0, level-1, and level-2. Higher levels shall be reflected by level-1 or level-2 because odd-level and even-level behaviors coincide. Table 2 indicates $d^*(k)$ by sampling method and jury size. Notably, the uninformative behavior of level-0 is not useful in minimizing errors and therefore $d^*(0) = 0$. Furthermore, the average level μ_d^* is relatively high, especially for $n = 6$, and translates in a relatively high ideal share of level-2.

[6] Alternatively, it can be assumed that every juror believes with some small probability $\epsilon > 0$ that every other juror votes randomly. This has no effect on the predictions of our model, but changes the upper bound for ϵ stated in Proposition 4.

	O		W	
$d^*(k)$	$n=3$	$n=6$	$n=3$	$n=6$
$k=0$	.00	.00	.00	.00
$k=1$	.50	.29	.68	.34
$k=2$	.50	.71	.32	.66
μ_d^*	1.50	1.71	1.32	1.66

Table 2: Optimal Level-k distribution $d^*(k)$ by treatment.

Table 3 shows the aggregate predictions $\sigma(s)$ following $d^*(k)$. For both O and W, levels of $\sigma(b)$ increase with jury size. Interestingly, these predictions coincide with the NE predictions, which also minimize the probability of jury error.[7] Intuitively, only level-2 jurors vote strategically red after a blue signal, which is required with some probability in order to correctly decide for the true state S under the unanimity rule. Table 4 gives the minimized probabilities of jury errors.

	O		W	
	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$
$n=3$	1	.50	1	.32
$n=6$	1	.71	1	.66

Table 3: Nash equilibrium and d^* predictions for $\sigma(s)$ by treatment.

Another prediction relates to the jury's accuracy to predict the true state of the world. The two types of errors are quantified in Table 4. Under Nash and $d^*(k)$ behaviour, these do not change with the jury size in O.

	O		W	
	$Pr(R\|B)$	$Pr(B\|R)$	$Pr(R\|B)$	$Pr(B\|R)$
$n=3$	.25	.50	.16	.54
$n=6$	.25	.50	.21	.52

Table 4: Probabilities of incorrect jury decisions in Nash equilibrium and under $d^*(k)$.

[7] Our W treatment predictions slightly differ from the W predictions of GMP who implement $p = \frac{7}{10}$ and not $p = \frac{2}{3}$. They have not considered the case without replacement.

3. Experimental design

In this study, we employ a 2×2 within-subject experimental design, varying the size of the jury, $n \in \{3, 6\}$, in one dimension and the sampling of the signal, without (O) or with replacement (W), in the other. This implies four treatments, 3O, 6O, 3W, 6W. Signals are balls drawn from a red or blue urn, $S \in \{R, B\}$, each of which is set to be equally likely to occur. Moreover, we assume the cost of either type of error to be equal ($q = \frac{1}{2}$) and the probability of drawing a ball of the same color as the urn to be twice more likely ($p = \frac{2}{3}$).

3.1. Team Communication

To ascertain subjects' level of thinking, we conduct the experiment with an intra-team communication protocol that yields incentivized written accounts of their individual reasoning (Burchardi and Penczynski, 2014).[8]

The communication protocol incentivizes the subjects' messages within their respective teams as follows. Subjects are randomly assigned into teams of two players. The two members are connected through the modified chat module of the experiment software. Once subjects know the decision problem, each team member can state a so-called "suggested decision" and justify it in a written message. After the suggested decision is made, the suggestions and messages get exchanged simultaneously. In a next step, both team members individually state their "final decision". For each decision, one of the two team members' final decisions is chosen randomly by the computer to count as the "team's action". This construction provides incentives to state the full reasoning underlying the suggested decision in a clear and convincing way. The message is entered in free form without explicit space or time limitations.

Note that each team takes the role of a single juror and members of the same team always observe one and the same signal. Juries are therefore composed of n teams and the jury decision derives from the n final decisions. However, our main analysis focuses on the suggested decisions and messages of the individual subjects.[9]

3.2. Experimental procedures

We conducted six experimental sessions in the Experimental Economics Laboratory at the University of Mannheim (mLab). A total of 96 subjects participated in the

[8] The experimental instructions are reprinted in section C of the online appendix. The experiments were programmed and conducted with the software z-Tree (Fischbacher, 2007).

[9] Arad et al. (2022) show that the team setup does not introduce systematic differences in strategic decisions compared to a setup with individuals.

experiments. Subjects were recruited from the University of Mannheim's subject pool. The sessions respectively had 12, 12, 24, 24, 12 and 12 subjects. All of the subjects were either Bachelor (69), Master (20) or Doctoral (7) students of the University of Mannheim. Subjects' mean and median ages were 22. 52% of the subjects were female. Out of 96 subjects, 27 were studying Economics, 11 of them first year, 6 of them second year, 6 of them third year and 4 of them forth year or above. 45% of the subjects had prior training in game theory and 95% of them had previously participated in laboratory experiments.

Each session began with an initial test phase, during which subjects familiarized themselves with the team communication procedures. During this test phase, subjects answered two unrelated questions involving guessing the date of two different historical events. Subsequently, subjects played the jury voting game in four consecutive parts, with each part corresponding to a different treatment. Each part consisted of two periods, with subjects playing the treatment-specific variation of the game in each period. For every period, subjects were randomly reassigned to a new team and a jury.[10] Prior to each part, subjects were provided with instructions relevant only to that specific treatment. At the end of each period, subjects were provided with the aggregated decision of their jury and the resulting payoff for that period. After completing a part, subjects received a new set of instructions for the subsequent one. The experimenter read the instructions aloud and addressed any clarification questions publicly.

The sessions were organized such that half of them followed an order in which subjects first played O (without replacement) treatments for jury sizes of three and six, respectively, followed by W (with replacement) treatments. The other half of the sessions featured a reversed order for the sampling treatments while maintaining the same order for the jury sizes.[11]

	O	W	$\sum$
$n = 3$	192	192	384
$n = 6$	192	168	360
$\sum$	384	360	744

Table 5: Number of observations by treatments.

In sum, 744 observations are collected in 6 sessions. Table 5 indicates 24 fewer

[10] Upon investigation, we have found no significant difference between the consecutive periods for each treatment. Consequently, we combined the data from both periods to increase the number of observations for our analysis.

[11] Order effects are analyzed and discussed in section 6.1.

observations for 6W treatment because of an imposed end in one session, in which several subjects took significantly more time to make their decisions than expected.[12]

Remuneration for subjects was structured as follows: for each period, subjects earned €2 if their jury reached a correct decision, and €0.20 if the decision was incorrect. A show-up fee of €4 was provided, with additional earnings based on the accuracy of their jury's decisions averaging at €8.9, and ranging from a minimum of €5 to a maximum of €12.4. The subjects received their payment after the experiment in private and in cash.

3.3. Classification Process

The messages are classified independently by two research assistants (RA). For each individual decision's message they indicated the level of reasoning that the message corresponds to most closely. For this task, the RAs are introduced to the level-k model and received detailed instructions about characteristics of the individual types.[13]

The following features of the levels of reasoning are derived from the model and guide the classification process (similar to Burchardi and Penczynski, 2014; Penczynski, 2017). Level-0 play corresponds to choosing randomly, entirely without justification or with some justification completely unrelated to the task. Level-1 jurors always follow their own signal. They may argue in favor of playing their own signal through some probability argument. Level-2 reasoning assumes that all other jury members follow their signal and suggests a way to best respond to that. Level-3 reasoning is aware that people best respond to a belief that others follow their signal by voting red. Since level-3's best response is to follow their signal, level-3 reasoning might have similarities to level-1 reasoning.

The classification procedure starts with both RAs providing independent sets of classifications. Then, both are anonymously informed about the classifications of the other RA and have the possibility to simultaneously revise their own classification. This revision process is repeated twice. These iterations allow them to reconsider diverging classifications and to screen errors or misperceptions.[14]

Table 6 indicates that out of 744 observations and hence message opportunities, 529 (71.1%) sent a message, which was classified in all except 4 cases. Out of the classified

[12] GMP faced a similar issue with their 6W treatment.

[13] Classification instructions for the RAs are reprinted in section D of the online appendix.

[14] The initial agreement rate was 77.6%. After the first revision it increased to 87.4%, and after the second revision, the final agreement rate was 93.2%. See Burchardi and Penczynski (2014), Eich and Penczynski (2016), and Penczynski (2019) for further evidence on the robustness and replicability of this kind of classification.

messages, 493 (93.2%) had a matching level classification by the RAs. Only observations with classified messages enter our analysis. Table 7 shows that the percentage of messages classified out of all observations is stable across treatments. Table 8 shows message examples that are classified for each type of level of thinking.

	Message Sent		No Message Sent
	Classified	Unclassified	
	525	4	215
Matched	493		

Table 6: Number of observations by messages and classification.

	O	W	$\sum$
$n = 3$	.69	.65	.67
$n = 6$	.65	.66	.66
$\sum$	.67	.66	.66

Table 7: Ratio of observations with classified messages by treatments.

Level	**Message**
L0	I don't have a clue.
L0	Did not exactly understand this experiment seems to be just depending on luck.
L1	The probability of our urn being the color of the ball is 2/3 while the probability of the other color is 1/3.
L1	Hm.. the chance of it being the correct color is higher than it being the wrong one.
L2	I'd take red, because if any other is taking blue, it'll be blue anyways.
L2	I suggest we go for red because our decision won't be decisive in the committee's vote if the others go for blue. We don't hurt anyone with this decision.
L3	I think we should stay at blue because the probability of the urn to be blue is 50/50. So the others may decide to take red as they assume that one team will choose blue but if every team thinks in this way we would lose.
L3	Risky to vote blue but others may not vote blue even when draw blue. I say we vote blue.

Table 8: Examples of Messages

3.3.1. GPT Classification

We have further leveraged GPT-4 in assessing the feasibility of classifications through Large Language Models (LLMs) and conducting robustness checks on the classification outcomes of the RAs. The process commenced with prompting GPT-4 using the classification guidelines prepared for the RAs.[15] Subsequently, we fed the messages to GPT-4 for classification. The classifications by GPT-4 aligned with 94.5% of the RAs' classifications.

4. Hypotheses

The literature has identified many classes of games, in which subjects apply level-k reasoning. We therefore expect that such reasoning is also used for jury voting (Crawford et al., 2013).

Hypothesis 1 *Jury member decisions are governed by level-k reasoning.*

Hypotheses 2 and 3 express some of the theoretical findings from section 2.4.

Hypothesis 2 *Given a level of reasoning, the behaviour will not differ by treatment. Hence, aggregate $\sigma(b)$ will not depend on the treatments, but on the level-k distribution $d(k)$ only.*

Hypothesis 3 *The jury error rates in terms of $Pr(R|B)$ and $Pr(B|R)$ are a function of the level-k distribution $d(k)$.*

What could influence the level-k distribution? At various points in the literature, the dependence of the level-k distribution on game and population characteristics has been discussed and empirically documented (Alaoui and Penta, 2016; Penczynski, 2016b, 2017; Koch and Penczynski, 2018).

Increasing task complexity raises the cognitive cost of strategic thinking, while reduced perceived pivotality curtails the motivation for such thought. In our experiment, we selected the sampling method and jury size as treatment dimensions, as each influences task complexity by changing signal realizations within the jury, which in turn affects a juror's perceived pivotality. A jury size of 3 presents fewer signal realizations than a jury of 6, irrespective of the sampling method. We hypothesize that such decreased complexity leaves more cognitive capacity for strategic deliberation and thus leads to a higher mean level of reasoning.

[15] See section B of the online appendix for the used prompt and the details of the GPT-4 classification procedure.

Hypothesis 4 *The smaller number of possible signal realizations in jury size 3 compared to jury size 6 frees cognitive capacity and leads to the observation of a higher average level of reasoning μ_d.*

In O, for a given state S, the distribution of red and blue signals within the jury is predetermined and hence known by the jury members. Rabin (2002) models the belief in the law of the small numbers by means of "without replacement" sampling, leading us to expect that O makes deliberation easier for subjects. In W, many more signal realizations are possible, especially with larger jury sizes n.

Hypothesis 5 *The smaller number of possible signal realizations in O compared to W frees cognitive capacity and leads to the observation of a higher average level of reasoning μ_d.*

5. Results

Section 3.3 has shown that the RAs agreed about the content of messages in 93.2% of the cases. Table 9 shows the aggregated level-k distribution according to these classifications. The distribution $d(k)$ is non-degenerate and features a heterogeneity of types, a hump-shape with mode behaviour at level-1, and hardly any level-3 behaviour. All of these are expected and common traits of level-k distributions as observed in other contexts (Crawford et al., 2013). An average level μ_d of 1.12 is well within the range between 1 and 1.5 that the literature commonly observes (Camerer et al., 2004; Costa-Gomes and Crawford, 2006; Burchardi and Penczynski, 2014).[16]

	$d(k)$
$k=0$	.21
$k=1$	.49
$k=2$	.29
$k=3$	.02
μ_d	1.12

Table 9: Aggregate level-k distribution $d(k)$.

[16] In each treatment, the balance between the received signals and the state of the urn is checked. The proportions of the red signals received were as follows: 0.48 for 3O, 0.54 for 6O, 0.53 for 3W, and 0.51 for 6W. Similarly, the proportions of the red states were found to be: 0.47 for 3O, 0.66 for 6O, 0.53 for 3W, and 0.49 for 6W. In sum, barring the imbalance of red and blue states for 6O, the treatments are balanced in terms of red and blue signals and states.

Result 1 *According to the message classification, jury member decisions are governed by level-k reasoning in a similar fashion as other strategic decisions in the literature.*

Table 10 shows behaviour $\sigma(s)$ for both signals by treatment and levels. While levels are inferred from messages without explicit knowledge of the suggested action, the behaviour of $\sigma(s)$ within levels is not statistically different across treatments (Fisher exact test, $p_{min} > 0.187$) except for level-0 behavior for between 6O and 6W (Fisher exact test, $p = 0.081$).

Some degree of informativeness in level-0 voting is identified for W treatments (Fisher exact test, 3W: $p = 0.134$, 6W: $p = 0.041$), while level-1 and level-2 voting closely align with predictions. Level-1 voting is primarily informative and significantly differs between signals in all treatments (Fisher exact test, $p_{max} < 0.001$). Level-2 voting is predominantly strategic and not significantly dependent on the signal (Fisher exact test, $p_{min} > 0.23$)

	O				W			
	$n=3$		$n=6$		$n=3$		$n=6$	
$d(k)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$
$k=0$	.75	.64	.71	.73	.40	.86	.38	.75
$k=1$	.03	.97	.09	.97	.09	.97	.04	.97
$k=2$	.92	1.00	.91	1.00	.92	1.00	1.00	1.00
$k=3$	.00	–	.00	–	.00	–	–	–

Table 10: $\sigma(s)$ per treatment and level.

Table 11 aggregates these numbers over levels. For $\sigma(r)$, Table 11a shows that our results (ÇP) are close to the NE predictions and $d^*(k)$ implication of $\sigma(r) = 1$. Yet, due to level-0 voting, we reject the null hypothesis that they are 0.99 or above for all treatments (one-sample binomial test: for 3O, 6O, 6W $p_{max} < 0.001$; for 6W $p = 0.071$). For $\sigma(b)$, Table 11b shows that the jury size has less influence than NE and $d^*(k)$ would predict, both in our data and in GMP's results. Specifically, in aggregate, $\sigma(b)$ is found not to be significantly different between jury sizes 3 and 6 for both O or W (Fisher exact test, O: $p = 0.371$; W: $p = 0.69$), and it is found to be significantly lower than predicted in 6O and 6W (one-sample proportion test , 6O: $p = 0.004$, 6W: $p < 0.001$).

In addition, note that in Table 11b, the aggregate strategic voting proportions, $\sigma(b)$, includes strategic votes by level-2 jurors but also non-strategic votes from level-0 jurors as well as mistakes of level-1 jurors. As a consequence, the $\sigma(b)$ proportions in Table 11b reflect more than purely strategic voting. When the non-strategic votes

are accounted for, the true aggregate strategic voting rates for 3O, 6O, 3W, and 6W decrease to .35, .34, .20, and .22 respectively. Under these adjusted values the $\sigma(b)$ proportions for 3O and 3W are also significantly lower than the NE predictions (one-sample proportion test, 3O: $p = 0.011$; 3W: $p = 0.037$).

	O		W		
	NE/d^*	ÇP	NE/d^*	ÇP	GMP
$n = 3$	1	.92	1	.97	.95
$n = 6$	1	.93	1	.90	.90

(a) $\sigma(r)$.

	O		W		
	NE/d^*	ÇP	NE/d^*	ÇP	GMP
$n = 3$	.50	.46	.32	.32	.36
$n = 6$	.71	.55	.66	.36	.48

(b) $\sigma(b)$.

Table 11: Nash equilibrium and $d^*(k)$ predictions and empirical results for behavior.

The logit regression in Table 12 summarizes these findings and shows on the one hand the different and significant impact of different levels on the voting behavior compared to level-0 – especially for $\sigma(b)$ – and on the other hand the minor and insignificant impact of the treatments.

Result 2 *Given a level of reasoning, behavior does not differ across treatments. Aggregate behavior $\sigma(b)$ depends less on the treatments than predicted by NE and $d^*(k)$.*

The observed behavior in terms of $\sigma(s)$ implies error rates as presented in Table 13.[17] For each type of error, the observed error rate deviates in the same direction from the prediction in all treatments. In Table 13a, the error rate $Pr(B|R)$ is significantly higher in 6 than in 3 for both O and W and consequently higher than under NE or d^* (Wilcoxon ranksum test[18], $p < 0.001$). Conversely, in Table 13b, the error rate

[17] Due to our intra-team methodology, half of the time, a decision made by a subject is not reflected in the final team decision. As a result, it is not possible to trivially sum the cases in which the jury's aggregated decision is correct using the experimental data of jury decisions. Instead, using the mean values for $\sigma(r)$ and $\sigma(b)$, for each treatment and state, we calculated the expected jury accuracies by plugging $\sigma(r)$ and $\sigma(b)$ into the formulas stated in Proposition 1 and Corollary 1.1 for ρ_R and ρ_B for W and O cases respectively. In GMP, accuracy values are derived directly from the jury decisions.

[18] Wilcoxon ranksum test is performed on the bootstrapped distributions for the error rates. These distributions are generated by calculating the error rates from the re-sampled with replacement vote variable.

Table 12: Logit regressions with average marginal effects on $\sigma(s)$.

	M_1		M_2	
	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$
Level-1	-.508***	.002***	-.502***	.002***
	(.072)	(.001)	(.070)	(.001)
Level-2	.503***	.090***	.518***	.093***
	(.123)	(.027)	(.118)	(.025)
Level-3	-.407***		-.407***	
	(.056)		(.055)	
$6O$	.062	-.000		
	(.149)	(.000)		
$3W$	-.049	.000		
	(.119)	(.000)		
$6W$	-.096	.000		
	(.101)	(.000)		

Notes: Values in parenthesis represent the standard errors clustered by subjects. '***' represents $p < 0.001$ significance. There are no level-3 observations for $\sigma(r)$ cases. M2 only considers the level variable while M1 additionally includes treatment variable. We compared the fit of two logistic regression models using a likelihood-ratio test, we fail to reject M2 over M1 for each subsample ($p_{min} > 0.52$). Hence, we did not find evidence that including the treatment factors in the model significantly improved the fit to the data.

$Pr(R|B)$ is significantly lower (Wilcoxon ranksum test, $p < 0.001$). Both movements can be explained with the lower than optimal sophistication in the observed $d(k)$ and the resulting infrequent strategic voting with $\sigma(b) = 1$. With the increased blue votes, the probability $Pr(B|R)$ – acquitting a guilty – increases and $Pr(R|B)$ – convicting an innocent – decreases.

Result 3 *In line with the observed level-k distribution, the jury accuracy deviates from the optimal accuracy in that convictions are less likely, independent of the state S.*

Table 14 shows the level distribution $d(k)$ by treatment. At first sight, the distributions and the mean levels μ_d are supportive of Hypotheses 5 and 4 that both treatment

	O		W		
	NE/d^*	ÇP	NE/d^*	ÇP	GMP
$n = 3$	.50	.61	.54	.57	.53
$n = 6$	.50	.78	.52	.86	.73

(a) $Pr(B|R)$.

	O		W		
	NE/d^*	ÇP	NE/d^*	ÇP	GMP
$n = 3$	.25	.19	.16	.16	.19
$n = 6$	.25	.08	.21	.02	.03

(b) $Pr(R|B)$.

Table 13: Nash equilibrium and $d^*(k)$ predictions, and empirical results for the error rates.

dimensions have an influence on the sophistication of strategic thinking.

	O		W	
$d(k)$	$n = 3$	$n = 6$	$n = 3$	$n = 6$
$k = 0$	.15	.24	.14	.32
$k = 1$	.45	.40	.58	.51
$k = 2$	.37	.35	.26	.16
$k = 3$	.03	.01	.02	.00
μ_d	1.28	1.13	1.14	0.84

Table 14: Level-k distribution $d(k)$ by treatment.

Pooling the level-k distribution across two jury sizes, Table 15a shows that the average level of reasoning is significantly higher under $n = 3$ compared to $n = 6$ (Wilcoxon ranksum test, $p < 0.001$).[19] Specifically, level-0 is higher while level-1, level-2, and level-3 are lower in $n = 6$ compared to $n = 3$.[20]

Result 4 *The treatments with jury size $n = 3$, which produce a smaller number of possible signal realizations in a jury, feature higher average levels of reasoning.*

Pooling the level-k distributions across different sampling methods, Table 15b shows that average level of reasoning is significantly higher under O compared to W (Wilcoxon

[19] Since the number of level-3 jurors are relatively low, we have additionally considered the same test on the subset that excludes the level-3 data, and still found a significant difference ($p = 0.003$)

[20] The observed differences for level-0 and level-3 are found to be significantly different (Fisher exact test, level-0: $p < 0.001$; level-3: $p = 0.02$), for level-1, it is found to be marginally different (Fisher exact test, $p = 0.149$), and for level-2, it is found not to be significantly different (Fisher exact test, $p = 0.315$)

$d(k)$	$n=3$	$n=6$
$k=0$	.14	.28
$k=1$	.52	.45
$k=2$	.32	.27
$k=3$	.03	.00
μ_d	1.23	1.00

(a) Jury size $n=3$ and $n=6$.

$d(k)$	O	W
$k=0$	.19	.23
$k=1$	.43	.55
$k=2$	.36	.22
$k=3$	.02	.01
μ_d	1.22	1.01

(b) Sampling O and W.

Table 15: Level distributions $d(k)$ by sampling and jury size.

ranksum test, $p < 0.001$).[21] Specifically, the level-1 fraction is higher and the level-2 fraction is lower in W compared to O.[22]

Result 5 *The sampling method O, which produces a smaller number of possible signal realizations in a jury, features higher average levels of reasoning.*

We have additionally compared the level distribution controlling for the signal. Although the average level of sophistication is observed to be higher for the blue signal, the difference between the two average levels of reasoning is found to only be marginally significant (Wilcoxon ranksum test, $p = 0.122$). As can be seen in Table 16, except for the level-3 ratios, the level of sophistication between the two distributions are quite close to each other [23]. Furthermore, when we exclude the few level-3 players from the data, the average strategic sophistication level for the blue signal subset becomes 1.11, and the marginal significance between the two subsets is lost (Wilcoxon ranksum test, $p = 0.318$). Upon receiving a red signal since there is no distinction in terms of voting behavior among levels, subjects potentially did not have the motivation to consider a higher level of reasoning. This might, in turn, have led to the lack of level-3 thinkers for the red signal cases, producing the observed marginal significant difference in the level of strategic sophistication between the two signals.

[21] Since the number of level-3 jurors are relatively low, we have additionally considered the same test on the subset that excludes the level-3 data, and still found a significant difference ($p = 0.002$).

[22] For level-1 and level-2, the observed differences are found to be significantly different (Fisher exact test, level-1: $p = 0.009$; level-2: $p < 0.001$) while for level-0 and level-3, they are found not to be significantly different (Fisher exact test, level-0: $p = 0.32$; level-3: $p = 0.29$).

[23] The observed differences for level-0, level-1 and level-2 are found not to be significantly different (Fisher exact test, $p_{max} > 0.24$), while for level-3, it is found to be significantly different (Fisher exact test, $p = 0.003$).

$d(k)$	$s = b$	$s = r$
$k = 0$	.20	.21
$k = 1$	.46	.51
$k = 2$	.31	.28
$k = 3$	.03	–
μ_d	1.17	1.08

Table 16: Distribution of Levels controlling for signal

6. Further explorations

6.1. Order effects

Considering our experiment involved four consecutive treatments, and the fact that the order of W and O treatments alternated between sessions, we explored potential fatigue and learning effects. These effects could be in action and might potentially confound our previously presented results.

6.1.1. Fatigue effect

As subjects advance through the treatments (rounds), cognitive exhaustion might set in, leading them to exhibit reduced strategic sophistication in later parts of the experiment. Therefore, if there is a noticeable fatigue effect, regardless of the specific treatments involved, one would anticipate a diminished average sophistication level in the later rounds of the experiment.

$d(k)$	H_1	H_2
$k = 0$	.25	.17
$k = 1$	.44	.52
$k = 2$	.3	.29
$k = 3$	.01	.02
μ_d	1.07	1.17

Table 17: Fatigue Effect

In Table 17, we categorized the data into two subsets based on treatments: those played during the first two rounds, labeled as H_1, and those from the last two rounds, labeled as H_2. We then analyzed the strategic level distribution across these subsets. Contrary to the expectation of a drop in sophistication due to potential fatigue, the results indicate a higher sophistication level in H_2 (Wilcoxon ranksum test, $p = 0.094$).

As a result, we deduce that fatigue did not have a predominant influence on the subjects, underscoring the robustness of our previously presented findings.

6.1.2. Learning effect

The learning effect, in contrast to the fatigue effect, might not manifest uniformly across treatments. Specifically, while the fatigue effect might consistently impact all the later rounds, the influence of the learning effect could differ based on the altering ordering of treatments.

When subjects first play O and then transition to W (O2W), an easy-to-hard learning effect might occur. The relatively lower complexity of the strategy space in O may enable subjects to better grasp the strategic nature of the game. This enhanced initial understanding of the game's strategic aspects could then foster greater strategic sophistication when subjects tackle more complex W that follow.

Conversely, when subjects start with W followed by O (W2O), a potential hard-to-easy learning effect might emerge. Initially engaging with W exposes subjects to a broader signal and strategy space, demanding heightened cognitive effort. As they transition to O, characterized by a strategy space with fewer possible outcomes, their prior experience in the complex dynamics of W could enhance their grasp of the game, leading to increased strategic sophistication compared to scenarios where O precede W.

Empirical evidence from domains such as motor skills (Wulf and Shea, 2002) and auditory skills (Liu et al., 2008; Church et al., 2013; Wisniewski et al., 2017) suggests that an easy-to-hard progression can enhance learning and performance compared to random or hard-to-easy orderings. In the domain of test-taking, some studies find no overall effect of question ordering on performance (Plake et al., 1982; Klimko, 1984), while others demonstrate a positive impact with easy-to-hard progression (Bassey et al., 2022; Hambleton and Traub, 1974). In contrast, hard-to-easy progression either shows no significant benefit (Hauck et al., 2017) or a negative effect on performance (Hambleton and Traub, 1974; Newman et al., 1988).

In Table 18, we do observe that jurors portray a significantly higher level of strategic sophistication in O2W relative to W2O (Wilcoxon ranksum test, $p = 0.016$). Given that O2W represents the learning effect via the easy-to-hard ordering while W2O represents the learning effect via the hard-to-easy ordering, this initial investigation hints at the possibility that only an easy-to-hard learning effect is present in our data.

Comparing the average level of strategic thinking between Tables 19a and 19b, we observe an easy-to-hard learning effect in W: the average level in O2W (1.13) is signif-

$d(k)$	O2W	W2O
$k=0$	.16	.24
$k=1$	.49	.48
$k=2$	.34	.26
$k=3$	.01	.02
μ_d	1.19	1.06

Table 18: Learning Effect

icantly higher than in W2O (0.89) (Wilcoxon ranksum test, $p = 0.003$). For O, however, the strategic thinking level in O2W (1.26) is higher than in W2O (1.19) (Wilcoxon ranksum test, $p = 0.1752$), which makes a hard-to-easy learning effect unplausible.

$d(k)$	O	W
$k=0$	.19	.14
$k=1$	.37	.59
$k=2$	.43	.26
$k=3$	.01	.01
μ_d	1.26	1.13

(a) O to W order, Ord_1

$d(k)$	O	W
$k=0$	.19	.31
$k=1$	.47	.5
$k=2$	.31	.18
$k=3$	.03	.01
μ_d	1.19	0.89

(b) W to O order, Ord_2

Table 19: Level distributions $d(k)$ by sampling and order.

Given the statistically significant influence of the easy-to-hard learning effect on W, we further examined whether our previously discussed results regarding the overall effect of sampling and group size remain robust when controlling for order. As illustrated in Table 19, the average sophistication level in O is consistently higher across both the O2W and W2O subsets. This disparity is statistically significant for both subsets (Wilcoxon ranksum test, O2W: $p = 0.053$; W2O: $p < 0.001$).

Similarly, Table 20 reveals that for both subsets, the average sophistication level is higher in the treatments with a smaller group size, and these differences are also significant (Wilcoxon ranksum test, O2W: $p = 0.102$; W2O: $p = 0.002$). Thus, our main results remain consistent after accounting for the ordering effect.

Given the documented presence of the easy-to-hard learning effect in our data, a natural inquiry arises: could this effect also influence performance across jury sizes of 3 and 6? It is worth noting that in our experimental sessions, we alternated only the order of sampling, while always maintaining the sequence of $n = 6$ treatments following $n = 3$ treatments. This consistent order means we lack the variation to explore this effect empirically. However, if the easy-to-hard learning effect is inflating the strategic sophistication in jury size 6 treatments, this implies an even more pronounced innate

$d(k)$	$n=3$	$n=6$
$k=0$	.12	.21
$k=1$	.51	.46
$k=2$	.35	.33
$k=3$	.02	–
μ_d	1.26	1.13

(a) O to W order

$d(k)$	$n=3$	$n=6$
$k=0$	.16	.34
$k=1$	.52	.44
$k=2$	.29	.22
$k=3$	.04	.01
μ_d	1.20	0.89

(b) W to O order

Table 20: Level distributions $d(k)$ by group size and order.

difference between the two jury sizes than reported. Hence, the possibility of such a learning effect would only further emphasize our findings on the differences in strategic thinking across the two jury sizes.

6.2. Influence of team communication on votes

In light of our intra-team communication experimental design, we sought to understand how communication impacts juror members' decisions after their interactions (see Penczynski, 2016a; Arad et al., 2022) with their teammates (partners). This inquiry is explored in Table 21.

Of the 493 messages, RAs were able to classify the strategic thinking levels of both the juror and her partner for 336 instances (68.2%). Of these, only 85 cases (25.3%) had differing pre-communication votes between the partners.

For partners that start with the same suggested vote, Table 21a shows that their subsequent decisions remain largely unchanged and unaffected by the level of strategic thinking conveyed with their communication. In contrast, for diverging suggested votes, Table 21b shows that the communicated message's strategic depth has a discernible influence. Specifically, jurors tend to change their vote in more instances when a higher level of reasoning is exhibited by their partner.

	Vote Changed?	
Higher Lv.	×	✓
Player	76	0
Neither	100	1
Partner	74	0

(a) Same Initial Vote.

	Vote Changed?	
Higher Lv.	×	✓
Player	23	9
Neither	13	6
Partner	16	18

(b) Different Initial Vote.

Table 21: Comparison of Vote Changes.

Note: Vote Change indicates whether the *Player*'s final vote is different from their suggested vote.

7. Concluding remarks

Our study provides evidence that strategic thinking is relevant in jury voting and appropriately modeled by the level-k model of reasoning. The predicted types and their behaviour align with the evaluation of written accounts and decisions observed in the experimental games. The model correctly predicts that – given a level of sophistication – behaviour is unresponsive to changes in jury size and signal sampling method. The experimental text and decision data support this prediction and show that behaviour reacts to treatments primarily because strategic sophistication responds to the cognitive complexity of the task at hand. Specifically, we find evidence that larger juries and the more involved "with replacement" sampling of signals lead to a lower strategic sophistication.

The deviations of the observed jury error rates from the minimal error rates in equilibrium can be viewed as a result of the sub-optimal distribution of the levels of reasoning in the subject sample. Interestingly, these error rates would be much higher, and jury votes under unanimity in large juries would basically be uninformative if only one type of level-k reasoning with informative or strategic voting was present. From that perspective, the heterogeneity in types being so close to the optimal distribution is what leads to the observed low error rate in the jury voting. So while many arguments in favor of jury diversity already exist, the study adds and exhibits in detail a stark reason for jury diversity in terms of strategic sophistication.

References

Agranov, Marina, Elizabeth Potamites, Andrew Schotter, and Chloe Tergiman, "Beliefs and endogenous cognitive levels: An experimental study," *Games and Economic Behavior*, 2012, *75* (2), 449–463.

Alaoui, Larbi and Antonio Penta, "Endogenous depth of reasoning," *The Review of Economic Studies*, 2016, *83* (4), 1297–1333.

Aljanabi, Mohammad, Mohanad Ghazi, Ahmed Hussein Ali, Saad Abas Abed et al., "ChatGPT: open possibilities," *Iraqi Journal For Computer Science and Mathematics*, 2023, *4* (1), 62–64.

Arad, Ayala, Kevin P. Grubiak, and Stefan P. Penczynski, "Does communicating within a team influence individuals' reasoning and decisions?," *Experimental Economics*, 2022, pp. 1–21.

Austen-Smith, David and Jeffrey S Banks, "Information aggregation, rationality, and the Condorcet jury theorem," *American political science review*, 1996, *90* (1), 34–45.

Bassey, Bassey Asuquo, Isaac Ubi, Effa German Anagbogu, and Valentine Joseph Owan, "Permutation of the UTME Multiple-Choice Test Items on Performance in Use of English and Mathematics among Prospective Higher Education Students," *The Journal of Social Sciences Research, ISSN (e)*, 2022, pp. 2411–9458.

Burchardi, Konrad B. and Stefan P. Penczynski, "Out of your mind: Eliciting individual reasoning in one shot games," *Games and Economic Behavior*, 2014, *84* (1), 39 – 57.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong, "A Cognitive Hierarchy Model of Games," *The Quarterly Journal of Economics*, August 2004, *119* (3), 861–898.

Church, Barbara A, Eduardo Mercado III, Matthew G Wisniewski, and Estella H Liu, "Temporal dynamics in auditory perceptual learning: impact of sequencing and incidental learning.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2013, *39* (1), 270.

Costa-Gomes, Miguel A. and Vincent P. Crawford, "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," *American Economic Review*, December 2006, *96* (5), 1737–1768.

Coughlan, Peter J, "In defense of unanimous jury verdicts: Mistrials, communication, and strategic voting," *American Political Science Review*, 2000, *94* (2), 375–393.

Crawford, Vincent P. and Nagore Iriberri, "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and

Overbidding in Private-Value Auctions?," *Econometrica*, November 2007, *75* (6), 1721–1770.

__ **, Miguel A. Costa-Gomes, and Nagore Iriberri**, "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications," *Journal of Economic Literature*, September 2013, *51* (1), 5–62.

Eich, Theresa and Stefan P. Penczynski, "On the replicability of intra-team communication classification," Working Paper, University of Mannheim 2016.

Eyster, Erik and Matthew Rabin, "Cursed Equilibrium," *Econometrica*, September 2005, *73* (5), 1623–1672.

Feddersen, Timothy and Wolfgang Pesendorfer, "Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting," *American Political Science Review*, 1998, pp. 23–35.

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, June 2007, *10* (2), 171–178.

Grimm, Veronika and Friederike Mengel, "Experiments on belief formation in networks," *Journal of the European Economic Association*, 2020, *18* (1), 49–82.

Guarnaschelli, Serena, Richard D McKelvey, and Thomas R Palfrey, "An experimental study of jury decision rules," *American Political Science Review*, 2000, pp. 407–423.

Hambleton, Ronald K and Ross E Traub, "The effects of item order on test performance and stress," *The Journal of Experimental Education*, 1974, *43* (1), 40–46.

Hauck, Kendall B, Maya A Mingo, and Robert L Williams, "A review of relationships between item sequence and performance on multiple-choice exams.," *Scholarship of Teaching and Learning in Psychology*, 2017, *3* (1), 58.

Klimko, Ivan P, "Item arrangement, cognitive entry characteristics, sex, and test anxiety as predictors of achievement examination performance," *The Journal of experimental education*, 1984, *52* (4), 214–219.

Koch, Christian and Stefan P Penczynski, "The winner's curse: Conditional reasoning and belief formation," *Journal of Economic Theory*, 2018, *174*, 57–102.

Li, Shengwu, "Obviously strategy-proof mechanisms," *American Economic Review*, 2017, *107* (11), 3257–3287.

Liu, Estella H, Eduardo Mercado III, Barbara A Church, and Itzel Orduña, "The easy-to-hard effect in human (Homo sapiens) and rat (Rattus norvegicus) auditory identification.," *Journal of Comparative Psychology*, 2008, *122* (2), 132.

McKelvey, Richard D. and Thomas R. Palfrey, "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, July 1995, *10* (1), 6–38.

Nagel, Rosemarie, "Unraveling in Guessing Games: An Experimental Study,"

American Economic Review, December 1995, *85* (5), 1313–1326.
Newman, Dianna L, Deborah K Kundert, David S Lane Jr, and Kay Sather Bull, "Effect of varying item order on multiple-choice test scores: Importance of statistical and cognitive difficulty," *Applied Measurement in education*, 1988, *1* (1), 89–97.
OpenAI, "GPT-4 technical report," *arXiv preprint*, 2023.
__ , "Prompt design,"
https://platform.openai.com/docs/guides/completion/prompt-design 2023.
Penczynski, Stefan P, "Persuasion: An experimental study of team decision making," *Journal of Economic Psychology*, 2016, *56*, 244–261.
Penczynski, Stefan P., "Strategic thinking: The influence of the game," *Journal of Economic Behavior & Organization*, 2016, *128*, 72–84.
Penczynski, Stefan P, "The nature of social learning: Experimental evidence," *European Economic Review*, 2017, *94*, 148–165.
__ , "Using machine learning for communication classification," *Experimental Economics*, 2019, *22* (4), 1002–1029.
Plake, Barbara S, Charles J Ansorge, Claire S Parker, and Steven R Lowry, "Effects of item arrangement, knowledge of arrangement test anxiety and sex on test performance," *Journal of Educational Measurement*, 1982, pp. 49–57.
Rabin, Matthew, "Inference by Believers in the Law of Small Numbers," *The Quarterly Journal of Economics*, 2002, *117* (3), 775–816.
Rapoport, Amnon, Darryl A. Seale, Ido Erev, and James A. Sundali, "Equilibrium Play in Large Group Market Entry Games," *Management Science*, 1998, *44* (1), 119–141.
Stahl, Dale O. and Paul W. Wilson, "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, July 1995, *10* (1), 218–254.
Tversky, Amos and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *science*, 1974, *185* (4157), 1124–1131.
Wisniewski, Matthew G, Milen L Radell, Barbara A Church, and Eduardo Mercado III, "Benefits of fading in perceptual learning are driven by more than dimensional attention," *Plos one*, 2017, *12* (7), e0180959.
Wulf, Gabriele and Charles H Shea, "Principles derived from the study of simple skills do not generalize to complex skill learning," *Psychonomic bulletin & review*, 2002, *9* (2), 185–211.

A. Proofs

A.1. Proof of Corollary 1.1

Following Feddersen and Pesendorfer (1998), we are looking for a responsive symmetric equilibrium in mixed strategy profiles. First note that a necessary condition for a mixed strategy profile is for a juror who receives a blue signal to be indifferent between voting red and voting blue. This occurs when the probability that the urn is red (conditional on the juror i's vote being pivotal and on her private signal to be blue) is equal to the threshold of reasonable doubt, q. Let $Pr(S|s, piv_i)$ represent the probability of state S, conditional on the signal s and on juror i being pivotal. Then a juror who receives a blue signal is indifferent between voting red and blue when

$$Pr(R|b, piv_i) = q. \tag{16}$$

Using Bayes formula we expand on equation (16) as:

$$\frac{Pr(piv_i|R,b)Pr(b|R)Pr(R)}{Pr(piv_i|R,b)Pr(b|R)Pr(R) + Pr(piv_i|B,b)Pr(b|B)Pr(B)} \tag{17}$$

where due to the without placement assumption we have:

$$Pr(piv_i|R,b) = \sigma(r)^{pn}\sigma(b)^{(1-p)n-1} \tag{18}$$

$$Pr(piv_i|B,b) = \sigma(r)^{(1-p)n}\sigma(b)^{pn-1} \tag{19}$$

In words, equation (18) describes the probability of being pivotal given n many balls and jurors where pn many balls are red and received by pn many other jurors, and $(1-p)n-1$ balls are blue and received by $(1-p)n-1$ many other jurors. Similarly, equation (19) describes the probability of being pivotal given n many balls and jurors where $(1-p)n$ many balls are red and received by $(1-p)n$ many other jurors, and $pn-1$ many balls are blue and received by $pn-1$ many other jurors. Note that in FP $Pr(piv_i|R,b)$ and $Pr(piv_i|B,b)$ are additively defined as $(\sigma(r)p + \sigma(b)(1-p))^{(n-1)}$ and $(\sigma(r)(1-p) + \sigma(b)p)^{(n-1)}$ respectively, while in the without replacement case, we have a ***simpler*** form as in equations (18) and (19). Using (18) and (19) in (17), we get:

$$Pr(R|b, piv_i) = \frac{(1-p)\sigma(b)^{(1-p)n-1}}{(1-p)\sigma(b)^{(1-p)n-1} + p\sigma(b)^{pn-1}} = q \tag{20}$$

Isolating $\sigma(b)$ in equation (20), we get:

$$\sigma(b) = \left(\frac{qp}{(1-p)(1-q)}\right)^{\frac{1}{n(1-2p)}} \tag{21}$$

Given $\sigma(b) > 0$, we have $Pr(R|b, piv_i) = q$. Clearly, $Pr(R|r, piv_i) > Pr(R|b, piv_i) = q$. Thus, we have $\sigma(r) = 1$.

The probabilities for the jury's decision to be wrong given the true state of the world are defined as $Pr(B|R)$ and $Pr(R|B)$. In FP, for the with replacement case, they are defined as $(\sigma(r)p + \sigma(b)(1-p))^n$ and $(\sigma(r)(1-p) + \sigma(b)p)^n$ respectively. In the without replacement case, due to the hypergeometric nature of the signals, they are defined as $Pr(B|R) = \sigma(r)^{pn}\sigma(b)^{(1-p)n}$ and $Pr(R|B) = \sigma(r)^{(1-p)n}\sigma(b)^{pn}$ in a similar fashion defined in equations 18 and 19.

Lastly, note that given (20), it can easily be shown that $\lim_{n\to\infty} \sigma(b) = 1$.

A.2. Proof of Proposition 2

For every juror i, her expected payoff for voting red given she receives a red signal is defined as follows:

$$\begin{aligned}
\mathbb{E}(u_i(\sigma_i(r) = 1) = \; & U(R,R)Pr(R|r)\alpha^r + U(R,B)Pr(B|r)\beta^r \\
& + U(B,B)Pr(B|r)(1-\beta^r) + U(B,R)Pr(R|r)(1-\alpha^r) \\
= \; & p(\pi\alpha^r - (1-q)(1-\alpha^r)) \\
& + (1-p)(\pi(1-\beta^r) - q\beta^r) && (22) \\
= \; & -p(1-q)(1-\alpha^r) - (1-p)q\beta^r && (23) \\
= \; & -(1-q)p + p(1-q)\alpha^r - (1-p)q\beta^r && (24)
\end{aligned}$$

Note that the step from equality (22) to (23) has been taken via the assumption $\pi = 0$ and the same step has also been taken in the rest of the derivations of this subsection.

Through similar steps, one can define the juror i's expected payoff for voting blue given she receives a red signal as:

$$\begin{aligned}
\mathbb{E}(u_i(\sigma_i(r) = 0) &= U(B,B)Pr(B|r) + U(B,R)Pr(R|r) && (25) \\
&= \pi(1-p) - (1-q)p \\
&= -(1-q)p && (26)
\end{aligned}$$

Notice that in equation (25) we do have a relatively simplified right hand side equation without the payoffs $U(R,R)$ and $U(R,B)$, and the pivotality probabilities. This is because since the juror votes blue, $U(R,R)$ and $U(R,B)$ cases never occurs and the belief on what the other jurors will do becomes irrelevant. Using equalities (24) and (26), we get the desired equality in (14) of the Proposition 2.

Next, for every juror i, we calculate her expected payoff for voting red given she

receives a blue signal as:

$$\begin{aligned}\mathbb{E}(u_i(\sigma_i(b)=1) = &= U(R,R)Pr(R|b)\alpha^b + U(R,B)Pr(B|b)\beta^b \\ &\quad + U(B,B)Pr(B|b)(1-\beta^b) + U(B,R)Pr(R|b)(1-\alpha^b) \\ &= \pi(1-p)\alpha^b - qp\beta^b + \pi p(1-\beta^b) - (1-q)(1-p)(1-\alpha^b) \\ &= -qp\beta^b - (1-q)(1-p)(1-\alpha^b) \\ &= -qp\beta^b + \alpha^b(1-q)(1-p) - (1-q)(1-p) \qquad (27)\end{aligned}$$

Through similar steps, one can define her expected payoff for voting blue given she receives a blue signal as:

$$\begin{aligned}\mathbb{E}(u_i(\sigma_i(b)=0) &= U(B,B)Pr(B|b) + U(B,R)Pr(R|b) \\ &= \pi p - (1-q)(1-p) \\ &= -(1-q)(1-p) \qquad (28)\end{aligned}$$

Using equalities (27) and (28), we get the desired equality in (15) of the Proposition 2.

A.3. Proof of Corollary 2.1

Using Proposition 2, under the additional simplifying assumption that $U(R,B) = U(B,R)$, we identify the strict inequality conditions for voting either red or blue (conditional on the signal) as in the Tables (22) and (23).

Note that for the inequalities in Tables (22) and (23) to be well defined, we assume, respectively, the conditions $\alpha^s + \beta^s > 0$ and $\alpha^s, \beta^s > 0$ to hold for $s \in \{r, b\}$. Either condition aims at avoiding the cases where a jury member is not pivotality in either state of the world. Table (22) offers a more general format for the conditions, and as a result, is presented in the main section of the paper. Table (23) provides a better starting point for various algebraic manipulations that takes places in the subsequent proofs that utilizes these boundary conditions.

	$Vote = r$	$Vote = b$
Signal = b	$\dfrac{\alpha^b}{\alpha^b + \beta^b} > p$	$\dfrac{\alpha^b}{\alpha^b + \beta^b} < p$
Signal = r	$\dfrac{\beta^r}{\alpha^r + \beta^r} < p$	$\dfrac{\beta^r}{\alpha^r + \beta^r} > p$

Table 22: Conditions for Informative and Strategic Voting

	$Vote = r$	$Vote = b$
Signal = b	$\frac{\alpha^b}{\beta^b} > \frac{p}{1-p}$	$\frac{\alpha^b}{\beta^b} < \frac{p}{1-p}$
Signal = r	$\frac{\alpha^r}{\beta^r} > \frac{1-p}{p}$	$\frac{\alpha^r}{\beta^r} < \frac{1-p}{p}$

Table 23: Conditions for Informative and Strategic Voting

In the with replacement case, since the signals are independent of each other, we have $\alpha^i = \alpha^j$ and $\beta^i = \beta^j$ for $i \neq j$ and $i, j \in \{r, b\}$. Hence, for ease of notation, we can drop the signal superscripts. Then, given p and n, we define the beliefs about pivotalities for each state as:

$$\alpha = (p\sigma(r) + (1-p)\sigma(b))^{n-1} \tag{29}$$

$$\beta = ((1-p)\sigma(r) + p\sigma(b))^{n-1} \tag{30}$$

In the without replacement case, given p and n, we define the beliefs on pivotalities for each state and signal received as:

Given signal is red

$$\alpha^r = \sigma(r)^{(pn-1)}\sigma(b)^{((1-p)n)} \tag{31}$$

$$\beta^r = \sigma(r)^{((1-p)n-1)}\sigma(b)^{(pn)} \tag{32}$$

Given signal is blue

$$\alpha^b = \sigma(r)^{(pn)}\sigma(b)^{((1-p)n-1)} \tag{33}$$

$$\beta^b = \sigma(r)^{((1-p)n)}\sigma(b)^{(pn-1)} \tag{34}$$

Using equations (31)-(34) and Table 23, we have that, in the without replacement case, a juror votes informatively if and only if $\left(\frac{1-p}{p}\right)^{\frac{1}{n(2p-1)}} < \frac{\sigma(r)}{\sigma(b)} < \left(\frac{p}{1-p}\right)^{\frac{1}{n(2p-1)}}$; and a juror votes strategically if and only if $\left(\frac{p}{1-p}\right)^{\frac{1}{n(2p-1)}} < \frac{\sigma(r)}{\sigma(b)}$.

A.4. Proof of Proposition 3

First of all, note that for the with replacement case, since the signals are independent of each other, we have $\alpha^r = \alpha^b$ and $\beta^r = \beta^b$. Henceforth, in the rest of the proof we will omit the signal subscripts for ease of notation.

Secondly, we denote the level of the juror in their pivotality probability as α_k and β_k for $k \in \mathbb{N}^0$.

Level-1 By assumption, a level-0 juror votes uninformatively. This translates to $\alpha_1 = \beta_1 > 0$ and $\frac{\alpha_1}{\beta_1} = 1$. By Table 23, a level-1 juror votes blue when the signal is blue, $\sigma_1(b) = 0$, if and only if $\frac{\alpha_1}{\beta_1} < \frac{p}{1-p}$, and votes red when the signal is red, $\sigma_1(r) = 1$, if and only if $\frac{\alpha_1}{\beta_1} > \frac{1-p}{p}$. Since $p > \frac{1}{2}$ and $\frac{\alpha_1}{\beta_1} = 1$ by assumption, the informative voting condition for both signals is satisfied and level-1 juror always votes informatively.

Level-2 By the above discussion, a level-1 juror always votes informatively. By definition, a level-1 juror receives the signal r with probability p in state R and with probability $1-p$ in state B. Hence, assuming the state is R and there are n jurors in total, a level-2 juror believes to be pivotal with probability $\alpha_2 = p^{n-1}$; and assuming the state is B, she believes to be pivotal with probability $\beta_2 = (1-p)^{n-1}$. Thus, we have $\frac{\alpha_2}{\beta_2} = \left(\frac{p}{1-p}\right)^{(n-1)}$. Given $p > \frac{1}{2}$, we have $\frac{p}{1-p} > 1 > \frac{1-p}{p}$. Moreover note that $\forall k > 1$, we have $\frac{p}{1-p} < \left(\frac{p}{1-p}\right)^k$. Setting $k = n-1$ and noting $k > 1$ is equivalent to $n > 2$, we have $\frac{1-p}{p} < \frac{p}{1-p} < \left(\frac{p}{1-p}\right)^{(n-1)}$. Hence both conditions for strategic voting in Table 23 are satisfied.

Level-3 Given a level-2 juror always votes red, a level-3 juror believes to be always pivotal, $\alpha_3 = \beta_3 = 1$. Given $\frac{\alpha_3}{\beta_3} = 1$, analogue to the level-1 behavior, every level-3 juror votes informatively.

Level-4 and above Due to the assumed degenerate population belief on the next lower level $k-1$, every even-leveled juror behaves the same way as a level-2 juror and always votes red. Furthermore, every odd-leveled juror behaves the same way as a level-3 juror behaves and always votes informatively.

A.5. Proof of Proposition 4

In the following proof, we denote the level of the juror and the received signal in their pivotality probability as α_k^s and β_k^s for $k \in \mathbb{N}^0$ and $s \in \{r, b\}$. If the signal subscript is omitted in the subsection of the proof, this indicates that we have $\alpha^r = \alpha^b$ and $\beta^r = \beta^b$.

Level-1 By assumption a level-0 juror votes uninformatively. Introducing an ϵ possibility to make a mistake in a symmetric manner does not change this fact. Hence, we have $\alpha_1 = \beta_1 > 0$ which, in turn, implies $\frac{\alpha_1}{\beta_1} = 1$. By Table (23), a level-1 juror votes blue when the signal is blue if and only if $\frac{\alpha_1^b}{\beta_1^b} < \frac{p}{1-p}$, and votes red when the signal

is red if and only if $\frac{\alpha_1^r}{\beta_1^r} > \frac{1-p}{p}$. Since $p > \frac{1}{2}$ and $\frac{\alpha_1}{\beta_1} = 1$ by assumption, informative voting condition for either type of signal received is always satisfied and a level-1 juror always votes informatively.

Level-2 Based on the above discussion, a level-1 juror always intend to vote informatively. Also note that she is assumed to make a mistake with some probability $\epsilon > 0$ and votes against her signal. Given a juror receives a red signal, her probability of being pivotal under states R and B are respectively defined as:

$$\alpha_2^r = (1-\epsilon)^{(np-1)}\epsilon^{(n(1-p))} \tag{35}$$

$$\beta_2^r = (1-\epsilon)^{n(1-p)-1}\epsilon^{(np)} \tag{36}$$

Using equation (35) and (36), we have $\frac{\alpha_2^r}{\beta_2^r} = \left(\frac{1-\epsilon}{\epsilon}\right)^{n(2p-1)}$. Using Table (23), given the received signal is red, in order for a level-2 juror to vote strategically we need:

$$\left(\frac{1-\epsilon}{\epsilon}\right)^{n(2p-1)} > \frac{(1-p)}{p} \tag{37}$$

With a bit of algebra, inequality (37) becomes:

$$\epsilon < \frac{1}{1+\left(\frac{(1-p)}{p}\right)^{\frac{1}{(2p-1)n}}} \tag{38}$$

Hence, given ϵ satisfies inequality (38), a level-2 juror votes red when she receives a red signal.

Next, assume that the level-2 juror receives a blue signal. Then we have:

$$\alpha_2^b = (1-\epsilon)^{(np)}\epsilon^{(n(1-p)-1)} \tag{39}$$

$$\beta_2^b = (1-\epsilon)^{n(1-p)}\epsilon^{(np-1)} \tag{40}$$

Using equation (39) and (40), we have $\frac{\alpha_2^b}{\beta_2^b} = \left(\frac{1-\epsilon}{\epsilon}\right)^{n(2p-1)}$. Using Table (23), given the received signal is blue, in order for a level-2 juror to vote strategically we need:

$$\left(\frac{1-\epsilon}{\epsilon}\right)^{n(2p-1)} > \frac{(p)}{1-p} \tag{41}$$

With a bit of algebra, equation (41) translates to:

$$\epsilon < \frac{1}{1+\left(\frac{p}{(1-p)}\right)^{\frac{1}{(2p-1)n}}} \tag{42}$$

Hence, given ϵ satisfies inequality (42), a level-2 juror votes red when she receives a blue signal.

Lastly, since $p > \frac{1}{2}$, we have:

$$\frac{1}{1+\left(\frac{p}{(1-p)}\right)^{\frac{1}{(2p-1)n}}} < \frac{1}{1+\left(\frac{(1-p)}{p}\right)^{\frac{1}{(2p-1)n}}}$$

Hence, inequality (42) is a sufficient condition for inequality (38) to be satisfied. Thus, given inequality (42) is satisfied, a level-2 juror always votes red.

Level-3 Given a level-2 juror always votes red, a level-3 juror believes to be pivotal with probability $\alpha_3 = \beta_3 = (1-\epsilon)^{(n-1)}$. Since $\frac{\alpha_3}{\beta_3} = 1$, just like a level-1 juror, every level-3 juror votes informatively.

Level-4 and above Given the cyclical nature of the behavior at even and odd levels of thinking, every even-leveled juror behaves the same way as a level-2 juror and always votes red.

Given the cyclical nature of the behavior at even and odd levels of thinking, every odd-leveled juror behaves the same way as a level-1 juror and always votes informatively.

A.6. Relaxing the level-0 jurors' uninformativeness assumption

Assume that a level-0 juror votes her signal with probability θ and votes the opposite of her signal with probability $1-\theta$. Given this assumption, a level-0 juror votes randomly given $\theta = \frac{1}{2}$ and as θ goes to 1 or 0, the informativeness of level-0 juror's vote increases. For $\theta = 1$, a level-0 juror is equivalent to a level-1 juror. Given this assumption, when the sampling is without replacement, we have the following pivotality values for the level-1 juror for the case where the signal is red as:

$$\alpha_1^r = \theta^{np-1}(1-\theta)^{n(1-p)} \tag{43}$$

$$\beta_1^r = \theta^{n(1-p)-1}(1-\theta)^{np} \tag{44}$$

Given equations 43 and 44, using the Table 23, we get the following inequality for level-1 juror to vote informatively when she receives a red signal:

$$\left(\frac{\theta}{1-\theta}\right)^{n(2p-1)} > \frac{1-p}{p} \tag{45}$$

Similarly for the case when the level-1 juror receives a blue signal, we have the following pivotality values:

$$\alpha_1^b = \theta^{np}(1-\theta)^{n(1-p)-1} \tag{46}$$

$$\beta_1^b = \theta^{n(1-p)}(1-\theta)^{np-1} \tag{47}$$

Given the equations 46 and 47, using the Table 23, we get the following inequality for level-1 juror to vote informatively when she receives a blue signal:

$$\left(\frac{\theta}{1-\theta}\right)^{n(2p-1)} > \frac{p}{1-p} \tag{48}$$

Define K as $\left(\frac{p}{1-p}\right)^{\frac{1}{n(2p-1)}}$. Then with a bit of algebra, we get the following condition for the level-1 juror to vote informatively:

$$\theta \in \left(\frac{1}{1+K}, \frac{K}{1+K}\right) \tag{49}$$

First note that given $p > \frac{1}{2}$ we have $\frac{1}{1+K} < \frac{K}{1+K}$. Then recall that θ represents the probability for a level-0 juror to vote informatively, and note that given the constraint for θ in (49), a level-1 juror votes informatively. Hence, given a level-0 juror votes her signal, i.e. votes informatively with probability θ such that it satisfies the constraint (49), a level-1 juror votes informatively.

Setting p to $\frac{2}{3}$, for $n = 3$, the inequality in (49) becomes $\left(\frac{1}{3}, \frac{2}{3}\right)$; and for $n = 6$, it becomes $\left(\frac{1}{1+\sqrt{2}}, \frac{\sqrt{2}}{1+\sqrt{2}}\right) \sim (0.41, 0.59)$.

For the with replacement case, first note that since the signals are independent, we have $\alpha_1^r = \alpha_1^b$ and $\beta_1^r = \beta_1^b$. Hence, we drop the superscript and define the following pivotality values:

$$\alpha_1 = (p\theta + (1-p)(1-\theta))^{n-1} \tag{50}$$

$$\beta_1 = ((1-p)\theta + p(1-\theta))^{n-1} \tag{51}$$

Given the equations in (50) and (51), using Table 23, we have the following inequality conditions for a level-1 juror to vote informatively:

$$\frac{p\theta + (1-p)(1-\theta)}{(1-p)\theta + p(1-\theta)} > \frac{1-p}{p} \tag{52}$$

$$\frac{p\theta + (1-p)(1-\theta)}{(1-p)\theta + p(1-\theta)} < \frac{p}{1-p} \tag{53}$$

Define $\widetilde{K}$ as $\left(\frac{1-p}{p}\right)^{\frac{1}{n-1}}$. Then with a bit of algebra, we get the following condition for a level-1 juror to vote informatively:

$$\theta \in \left(\frac{p(1+\widetilde{K}) - 1}{(2p-1)(1+\widetilde{K})}, \frac{p + \widetilde{K}(p-1)}{(2p-1)(1+\widetilde{K})}\right) \tag{54}$$

Setting p to $\frac{2}{3}$, for $n = 3$, the inequalities in (52) and (53) approximately translate to $\theta \in (0.24, 0.76)$ and for $n = 6$, they translate to $\theta \in (0.4, 0.6)$.

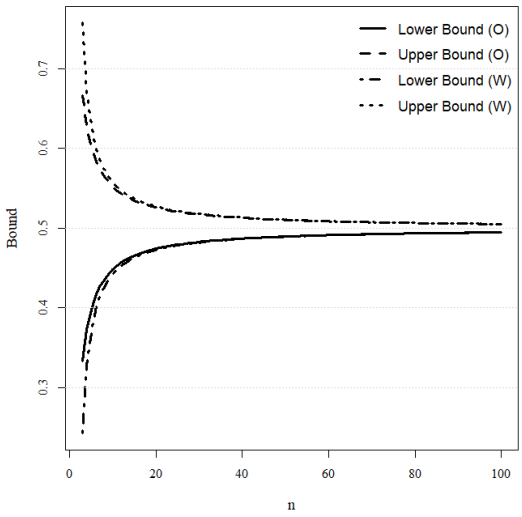

Figure 1: Bounds for θ given n and sampling method

As can be observed in Figure 1, for both sampling cases, as the group size increases, the range for θ, or in other words, the degree of a level-0 juror's informativeness in her vote decreases. For a level-1 juror to not vote informatively, i.e. to vote strategically, her belief on the informativeness of the aggregation of the other jurors' vote should overweight the informativeness of her signal. As the number of other jurors in the jury increases the necessary informativeness each of these jurors should provide with their vote for their aggregate informativeness to overweight the informativeness of the juror's signal decreases.

B. GPT-4 Classification Procedure

We provided GPT-4 with the following initial prompt, following the guidelines provided by OpenAI (OpenAI, 2023b):

> The total length of the content that I want to send you is too large to send in only one piece.
>
> To send you this content, I will follow these rules:
>
> [START PART 1/3] this is the content of the part 1 out of 3 in total [END PART 1/3]
>
> Then you just answer: "Received part 1/3"
>
> Once all parts are sent, I will indicate it by stating that 'ALL PARTS ARE SENT'.

> The text that I will provide you in parts is a classification instructions. Once all the parts of these instructions are sent, you will have an understanding on how to classify a text. Then I will provide a list of text for you to classify based on these instructions.

Following this initial prompt, we divided the classification instructions into three separate parts due to the prompt token limit of GPT-4 (OpenAI, 2023a). After providing the instructions, we presented approximately 30 texts (subject messages) to GPT-4 for classification. We repeated this process of classification until all the text data was classified.

Outputs for GPT-4 were obtained using ChatGPT version 4, updated last on March 24, 2023. We opted for ChatGPT to access GPT-4 as at the time of conducting our analysis there was no public access to GPT-4 via any other means. GPT-4 within ChatGPT is set to operate at an immutable temperature of 0.7 (Aljanabi et al., 2023). The temperature parameter is crucial in GPT models as it modulates output variability, with a setting of 0 providing uniform, consistent responses to identical prompts, and a setting of 1 introducing the greatest variation in responses (OpenAI, 2023a). Therefore, employing ChatGPT with its inherent output variability at a temperature of 0.7 introduces a notable level of unpredictability in classification tasks. To mitigate this and enhance the reliability of our classification, we duplicated the procedure, accepting classifications only when they aligned in both runs.

C. Experiment instructions

Introduction

You are about to participate in an experiment in team decision making. Please follow the instructions carefully.

In the experiment you may earn a considerable amount of money. Your decisions and the decisions of the other participants determine the amount. You will be instructed in detail how your earnings depend on your and the others' decisions. All that you earn is yours to keep, and will be paid to you in private and in cash, after today's session.

It is important to us that you remain silent and do not look at other people's screens. If you have any questions or need assistance of any kind, please raise your hand, and an experimenter will come to you. If you talk, exclaim out loud, etc., you will be asked to leave. Thank you.

Since this is a team experiment, you will at various times be matched randomly with another participant in this room in order to form a team that plays as a single entity. Your team's earnings will always be shared equally between you and your team partner.

The experiment consists of four parts (**Parts I, II, III and IV**). The parts are independent of each other but feature the same task in different settings. Each part consists of two rounds that require you to take a single decision. The way you interact as a team to take decisions will be the same throughout the experiment. Common features to the Parts will be given in the general instructions section.

Now, let us explain how your ***Team's Decision*** is determined. First of all, both you and your team partner will individually submit a ***Final Decision*** and the computer will randomly choose one of these two final decisions as your team's decision. The probability that your team partner's final decision is chosen is equal to the probability that your final decision will be chosen (i.e. your chances are 50:50). However, you have the possibility to influence your partner's final decision in the following way: Before you enter your final decision, you can propose to your partner a ***Suggested Decision*** and send her one and only one text ***Message***. *Note that this message is your only chance to convince your partner of the reasoning behind your suggested decision. Therefore, use the message to explain your suggested decision to your team partner.* After you finish entering your suggested decision and your message, these will be shown to your team partner. She will then make her final decision. Similarly, you will receive your partner's suggested decision and message. You will then also make your final decision. As indicated above, once you both enter your final decision, the computer chooses randomly one of your final decisions as your team's decision.

If you have any questions at this point, please raise your hand. In order for you to get familiar with the messaging system, you will now try it out in a **Test Period**. Please turn the page for further instructions.

Test period

A participant in this room is now randomly chosen to be your team partner. The **Test Period** has two rounds, with one question to answer in each round. Since this is only a test, your earnings will not depend on any decision taken now. In both test rounds, you will need to answer a question about the year of an historic event. The team that is closest to the correct year wins. Ties will be broken randomly by the computer.

As described, you will be able to send one ***Suggested Decision*** with your proposed year and an explaining ***Message***. After having read your partner's suggested decision and message, you will enter your ***Final Decision***. As described earlier, either your or your partner's final decision will be chosen randomly to be your ***Team's Decision***.

The messenger allows ***Messages*** of any size. However, you have to enter the message line by line since the input space is only one line. Within this line, you can delete text by using the usual "Backspace" button of your keyboard. By pressing "Enter" on the keyboard, you add the written sentence to the message. Please note that only added sentences will be sent and seen by your partner. *The words in the blue input line will* ***not*** *be sent*. You can always delete previously added sentences by clicking the "Clear Input" button. The number of lines you send is not limited. You can therefore send messages of any length. You finally send the message to your partner by clicking the "Send Message" button.

When you are ready, please click the "Ready" button to start the **Test Period**.

General Instructions

In every round of every part of the experiment you will be matched with a single, randomly chosen, different team partner. Together with other teams, which will also differ in every round, your team will face the following situation.

Your team along with other teams will constitute a ***committee*** in which each team has the right to a ***single vote***.

There will be *two* colored urns containing *some number* of colored balls. The urns will either be ***red or blue***, and either colored urn contains some number of red balls and some number of blue balls.

The color of the urn will determine the ratio of the number of red and blue balls in it. ***In the red urn***, there will be ***two times more red balls*** than blue balls and ***in the blue urn***, there will be ***two times more blue balls*** than red balls.

Every round, your newly formed committee will be assigned to a ***single urn*** whose ***color*** will be ***randomly determined*** by the computer as either blue or red ***with the equal probability***.

Your team – like all other teams in your committee – will observe the color of only one ball drawn from the urn assigned to your committee. Your task as a committee is to ***correctly guess the color of the urn***. The guess of the committee is determined by the votes of its teams.

The votes of the teams in the committee will be aggregated to a committee decision according to the ***unanimity voting rule***. The rules of the voting rule are as follows:

- If all the teams vote for the red color, then the color red will be the committee's guess.
- If at least one of the three teams votes for the blue color, then the color blue will be the committee's guess.

In other words, in order for the committee to select the red color as the guess, all the teams have to vote red. Thus, the teams need to *reach a consensus* in order to guess red; otherwise, the guess of the group will be blue.

If the committee correctly guesses the color of the urn they are assigned to, ***every team member*** in the committee will receive ***200 Eurocents***. If the committee does not correctly guess the color of the urn ,then ***every team member*** in the committee will receive ***20 Eurocents***.

Upon the observation of the color of your team's ball, you will send your team partner a ***Suggested Decision*** and a ***Message***. Remember to explain in the message your reasoning

behind your suggested decision. After this information is exchanged, both of you enter your ***Final Decision***, from which the computer randomly chooses the ***Team's Decision***.

The instructions of the Parts will specify the number of teams in the committee, the number of balls in the urn and the exact procedures of drawing a ball from the urn. Are there any questions at this point?

Part I

You are about to start Part I of the experiment. In each of the two rounds you will be matched with a new team partner and a new committee. Your team belongs to a committee that consists of three teams (your team and two other teams).

In this part, the balls are drawn from an urn that contains **two balls** of the same color of the urn and **one ball** of the "opposite" color. Your team will only observe the color of a single ball drawn from the urn. For all three teams, the balls will be drawn **without replacement**. That means that a drawn ball is not returned back to the urn for subsequent draws. There will therefore always be two teams observing the correct color and one team observing the incorrect color (with respect to the color of the urn assigned the teams). Your team will ***not*** know or observe the colors of the balls given to the other teams in your committee.

Upon the observation of the color of your team's ball, you will send your team partner a ***Suggested Decision*** and a ***Message***. Remember to explain in the message your reasoning behind your suggested decision. *(And note again that the words in the blue input line will **not** be sent. Press "Enter" to add them to the message.)* After this information is exchanged, both of you enter your ***Final Decision***, from which the computer randomly chooses the ***Team's Decision***.

When you click the "Ready" button, you will start **Part I** of the experiment.

Part II

You are about to start Part II of the experiment. In each of the two rounds you will be matched with a new team partner and a new committee. Your team belongs to a committee that consists of six teams (your team and five other teams).

In this part, the balls are drawn from an urn that contains **four balls** of the same color of the urn and **two balls** of the "opposite" color. Your team will only observe the color of a single ball drawn from the urn. For all six teams, the balls will be drawn **without replacement**. That means that a drawn ball is not returned back to the urn for subsequent draws. There will therefore always be four teams observing the correct color and two teams observing the incorrect color (with respect to the color of the urn assigned the teams). Your team will ***not*** know or observe the colors of the balls given to the other teams in your committee.

Upon the observation of the color of your team's ball, you will send your team partner a ***Suggested Decision*** and a ***Message***. Remember to explain in the message your reasoning behind your suggested decision. *(And note again that the words in the blue input line will **not** be sent. Press "Enter" to add them to the message.)* After this information is exchanged, both of you enter your ***Final Decision***, from which the computer randomly chooses the ***Team's Decision***.

When you click the "Ready" button, you will start **Part II** of the experiment.

Part III

You are about to start Part III of the experiment. In each of the two rounds you will be matched with a new team partner and a new committee. Your team belongs to a committee that consists of three teams (your team and two other teams).

In this part, the balls are drawn from an urn that contains **two balls** of the same color of the urn and **one ball** of the ”opposite” color. Your team will only observe the color of a single ball drawn from the urn. For all three teams, the balls will be drawn **with replacement**. That means that a drawn ball is returned to the urn for the subsequent draws. Independently of other teams' draws, each team will have a 2/3 chance of observing the correct color and a 1/3 chance of observing the incorrect color (with respect to the color of the urn assigned the teams). Your team will ***not*** know or observe the colors of the balls given to the other teams in your committee.

Upon the observation of the color of your team's ball, you will send your team partner a ***Suggested Decision*** and a ***Message***. Remember to explain in the message your reasoning behind your suggested decision. *(And note again that the words in the blue input line will **not** be sent. Press “Enter” to add them to the message.)* After this information is exchanged, both of you enter your ***Final Decision***, from which the computer randomly chooses the ***Team's Decision***.

When you click the “Ready” button, you will start **Part III** of the experiment.

Part IV

You are about to start Part IV of the experiment. In each of the two rounds you will be matched with a new team partner and a new committee. Your team belongs to a committee that consists of six teams (your team and five other teams).

In this part, the balls are drawn from an urn that contains **four balls** of the same color of the urn and **two balls** of the ”opposite” color. Your team will only observe the color of a single ball drawn from the urn. For all three teams, the balls will be drawn **with replacement**. That means that a drawn ball is returned to the urn for the subsequent draws. Independently of other teams' draws, each team will have a 2/3 chance of observing the correct color and a 1/3 chance of observing the incorrect color (with respect to the color of the urn assigned the teams). Your team will ***not*** know or observe the colors of the balls given to the other teams in your committee.

Upon the observation of the color of your team's ball, you will send your team partner a ***Suggested Decision*** and a ***Message***. Remember to explain in the message your reasoning behind your suggested decision. *(And note again that the words in the blue input line will **not** be sent. Press “Enter” to add them to the message.)* After this information is exchanged, both of you enter your ***Final Decision***, from which the computer randomly chooses the ***Team's Decision***.

When you click the “Ready” button, you will start **Part IV** of the experiment.

D. Classification instructions

D.1. Welcome

Thank you for participating in this experiment. In this document you will find instructions as to how this experiment works. You will be asked to classify messages that have been collected in an experiment on voting games. You as well are in an experiment which allows us to give you particular incentives and makes it easier for us to pay you.

To take part in the experiment, we assume that you are familiar with the level-k model as it has been introduced by Nagel (1995). However, in order to clarify potential questions of terminology, we reproduce the main features of the level-k model. In addition we provide detailed instructions of the original experiment, which explain the voting game and also give you a short introduction to voting games. Please read all information carefully.

Classification Task

General Comments

Your task is to classify the messages sent by subjects to their team member into their according level of thinking. You will read each comment and classify it according to the guidelines below. You can enter your assessment into the excel sheet provided to you. The excel sheet will have 6 different columns: The first four columns (which are already provided/filled) will identify the message by indicating the experiment, the part, the subject, the period. The fifth column is the classification column that we want you to fill and the sixth column is to indicate your personal comments on your task to further clarify your classification choice if necessary. The order of these columns will follow the transcript we will provide you.

It is very important that you double check whether the first 3 columns are filled correctly, i.e. that you enter the data for the correct subject, period, part and experiment. Based upon the guidelines below your task will be to fill up the classification column with an integer between 0 and 3 ("0" for level-0, "1" for level-1, "2" for level-2 and "3" for level-3) or leave it empty. If you find interesting elements that occur frequently but that have not been picked up by us, feel free to add a new column and mark all messages that contain the element. You can then specify to us in an email what exactly this element is.

For each individual classification, your assessment will be benchmarked against another classifier's assessment. Your personal remuneration is based on the number of matches of the level classification. A match is a classification that is congruent with the classification of another independent classifier. Each match will be remunerated with 0.07 Euro.

Please read this document and the instructions for the experiment entirely in order to get an overview and only then start the classification based on the player's sent message and action proposed. If you have any questions please do not hesitate to contact us.

The Original Experiment

A single experimental session consists of 4 parts and every part consists of 2 periods. In every part, every subject is randomly paired with another one to form a team. Depending on the treatment, every team then is randomly matched with 2 or 5 other teams to form a voting group. Then every group will be assigned to a blue or red urn and every team in the group will draw a ball from their assigned urn. The blue (red) urn contains twice more blue (red) balls then red (blue) balls. One treatment variable is the group size, in every experiment, in parts 1

(2) and 3 (4) the groups will consists of 3 (6) teams and accordingly the urns will contain 3 (6) balls. The other treatment variable is the draw mechanism, depending on the session, either for the first 2 parts or the last 2 parts, the balls will be drawn from the urn without replacement (i.e. any ball picked will not be placed back to the urn). For the other 2 parts, the balls will be drawn with replacement (i.e. any ball picked will be placed back to the urn before the next draw). In addition, we have included classroom experiments to the data set (experiments 7 and 8). In both experiments, the balls are drawn from the urn without replacement. Both experiments have two parts where in the first part the group size is 3 while in the second part the group size 6.

The goal of every group is to correctly state the color of the urn they are assigned to. Every team submits a vote (either red or blue) and if every team in the group votes for red then the group's decision will be red, otherwise it will be blue. Beyond this brief description of the experimental setup, please read the instruction sheets given to the subjects to have a better understanding of the situation of subjects [24].

How do the messages get produced? Every subject in a team observes the team's drawn ball and then makes a decision on the color of the urn. Then, they send a message to their team member explaining why they should vote for the suggested color. Next, the team members receive the suggestions and the message of their teammates, and make their final decision on their own. Either of the team member's decision is equally likely to be chosen as the team's final voting decision. This experimental methodology is developed to elicit subjects' reasoning through their sent messaged (see Burchardi and Penczynski, 2014). In the classroom experiments (experiments 7 and 8), subjects were not paired into teams but instead they were asked to elaborate on their reasoning for taking the action they have taken.

Model

General Model

It is assumed that you are familiar with the level-k model as it has been introduced by Nagel (1995) or represented by Camerer (2004). In order to clarify potential questions of terminology and introduce the main features of the model we quickly reproduce the main features of the model in the terminology used in this document. The level-k model of bounded rationality assumes that players only think through a certain number (k) of best responses. The model has four main ingredients:

Population distribution: This distribution reflects the proportion of types with a certain level $k \in N_0 = \{0, 1, 2, 3, 4, 5, \ldots\}$.

Level-0 distribution: By definition, a level-0 player does not best respond. Hence, his actions are random to the game and distributed randomly over the action space. In our case, the action space is $\mathcal{A} = \{Red, Blue\}$ where Red and $Blue$ represent the voting choice of the player. Note that, our model does not incorporate salience by assuming higher probabilities in the level-0 distributions for the action that is salient due the signal received (i.e. if a blue ball is received and the player is playing random, the player, due to the availability of the blue signal, will chose to vote blue)

Level-0 belief: In the model, the best responses of players with level $k > 0$ are anchored in what they believe the level-0 players play. Their level-0 belief might not be consistent with the level-0 distribution. For best responding, all that matters is the expected payoff from choosing an action from the action space $\mathcal{A} = \{Red, Blue\}$.

Population belief: Players do not expect other players to be of the same or a higher level of reasoning. For a level-k player, the population belief is therefore defined on the set of levels

[24] there are two versions of the instruction sheet where the versions vary in their of treatments. We are providing you with one copy. Please do not base the treatment ordering on the ordering of the instruction sheet.

strictly below k. It follows that level-0 players have no defined belief, level-1 players have a trivial belief with full probability mass on $\{0\}$ (i.e. the belief that everyone else is level-0), level-2 players have a well defined belief on $\{\{0\},\{1\}\}$. From level 3 higher order beliefs are relevant as level-3 players have to form a belief about level-2's beliefs.

Specific Model

We consider a game with n players (jurors in the voting context). The game starts by nature choosing a state of the world S in $\Omega = \{Red, Blue\}$ with probability r and $1-r$ respectively. The players do not observe the state, but each acquires a private signal s about the realized state of the world. If the true state is $Blue$, then each player observes an independent (or geometrically dependent[25]) Bernoulli random variable (the private signal) which is $blue$ with probability p and red with probability $1-p$ (and conversely for when the true state is Red). After observing their private signals, players chose an action a (a vote) from the action space $\mathcal{A} = \{Red, Blue\}$. Given the votes of the players, $1 \leq k* \leq n$ represents the number of votes needed for Red to be chosen for the aggregate decision. In other words, if $k*$ many or more players vote for Red, then the group decision is Red otherwise it is $Blue$. The utility of jury j when she takes action a with certainty given her signal is s and given the state is S is defined as $u_j(\sigma_j(s) = a, S)$. Given any signal s, the utility $u : A$ x $\Omega \mapsto \mathcal{R}$ for jury j is further defined by $u_j(\sigma_j(s) = blue, Blue) = u_j(\sigma_j(s) = red, Red) = 0$, $u_j(\sigma_j(s) = red, Blue) = -q$ and $u_j(\sigma_j(s) = blue, Red) = -(1-q)$, where $0 < q < 1$.

In all our experiments, we used $k* = n$ (i.e., the unanimity voting rule), $r = 0.5$, $q = 0.5$ and $p = \frac{2}{3}$. As previously explained treatments vary in the number of players n between 3 and 6.

Under the experimental set-up, the model provides the prototypical behavior of the subjects given their level as follows:

$n = 3$ ***or*** $n = 6$ ***and with replacement:***

When the blue ball is observed, optimal strategy for:
- level-1 player: vote blue
- level-2 player: vote red

When the red ball is observed, optimal strategy for:
- level-1 player: vote red
- level-2 player: vote red

$n = 3$ ***and without replacement:***

When the blue ball is observed, optimal strategy for:
- level-1 player: vote blue
- level-2 player: vote red

When the red ball is observed, optimal strategy for:
- level-1 player: vote red
- level-2 player: vote red or blue

(Voting red is strictly preferred to voting blue under additional assumption that the level-2 player assumes with some small probability ϵ that the other players are level-0

[25] Independent case refers to the aforementioned with replacement draws case while the geometrically dependent case refers to the without replacement case

$n = 6$ ***and without replacement:***

When the blue ball is observed, optimal strategy for:
- level-1 player: vote blue
- level-2 player: vote red or blue[26]

(Voting blue is strictly preferred to voting blue under additional assumption that the level-2 player assumes with some small probability ϵ that the other players are level-0 and not level-1)

When the red ball is observed, optimal strategy for:
- level-1 player: vote red
- level-2 player: vote red or blue

(Voting red is strictly preferred to voting blue under additional assumption that the level-2 player assumes with some small probability ϵ that the other players are level-0)

Guidelines for classification

General Comments:

- Subjects do not necessarily describe every step of their thinking; therefore, it may not always be obvious to decide which level they are. In many comments, any indications of a level of thinking may be partial or implicit, you should then indicate the most likely level of reasoning of the player.
- If the message indicates to simply refer to a previous message ('same as before/above'), then you can use the previous message's evaluation to determine the level of the current message. Please indicate this inference with a 1 in the column "Other message inference".
- If you are unsure of the level of the message, you should indicate the level you think is more likely.
- We have deliberately chosen not to disclose the action taken by the subject. You may still see in their comment which action they chose. We do not want you to base your classification on the action taken as it may be misleading.

Empty classification:

If no message has been formulated you should leave the classification empty. Also, you should leave the classification empty, if you are not sufficiently certain that any of the types below is capturing the strategic thinking in the message.

Level-0 Player:

Characteristics Chooses randomly, without justification or through some justification completely unrelated to the task. Might not have understood the game or shows no interest in the game or in thinking about it.

Examples 'My favorite color is blue, So I chose blue.'
'did not exactly understand this experiment seems to be just depending on luck'
'Just a guess'

[26]Note that for $n = 3$ voting red is strictly preferred to voting blue

Note Comments such as “It is obviously blueâ or âPlay red, trust me!â should not be considered as level-0 thinking as these comments to some extent signal some level of understanding/interest of the task. Such comments are likely to be level-1 comments yet without any additional information, you should leave the specific cell empty.

Level-1 Player:

Characteristics Always follows his own signal. The subject may argue in favor of playing his own signal through some probability argument

Examples 'The probability that the red ball we observe is out of the red urn is twice the probability that it is out of the blue urn'
'Our signal is blue. Letâs play blue.'

Note The key idea in defining a level-1 player is to identify some thinking process that signals the subjectâs interest/understanding of the task and the private signal. Furthermore, it is important that the subject does **not** offer any argument acknowledging the potential votes of the other teams and how to vote accordingly (i.e. adjusting the strategy given what others are expected to do).

Level-2 Player:

Characteristics Assume that all other players almost always follow their signal (i.e. she assumes almost all the other players are level-1 while an epsilon portion of them are level-0). Player does offer an argument acknowledging the potential votes of the other teams and how to vote accordingly (i.e. a best response given others are most likely playing their signal). In other words, if you identify any comment that indicates that the subject assumes (or considers the case) where the other players in her group play their signal, you should consider the possibility that the subject is a level-2 player.

Examples 'I have a blue ball. If we have the blue urn, someone else also has a blue ball and as a result our group will chose blue regardless of my vote. If we have the red urn, I am the only one with the blue ball and if I vote blue, we will chose the wrong urn. So I should vote for red.'
'In case two teams choose red and one chooses blue, blue will be taken. That means that choosing red has a higher chance of being a good decision.'
'I guess this is more about luck because there is no way to know it for sure. I would say blue just because of the higher probability. Also I like turtles Also it is likely that one other team will pick blue and then it is that color anyways'
'There is no point for us to take blue I think the chances for us to get the right color are higher if we stick with red' [red ball is observed]
'I suggest red because we donât hurt anyone with this decision If the others go for blue because they have a blue ball, the committeeâs decision will be blue regardless of our decision'
'[â¦] we could be the deciding vote for blue if the other two choose red' 'Choosing blue isnât as helpful as choosing red, because: only one blue ball can overturn our whole decision but only a unanimous decision for red can help us the same way'

Note In order to discern the two types, you should look for more than any trivial arguments such as the ones given under level-1. There may be cases where the message starts as a level-1 argument and then as the subjects elaborates on her reasoning, she starts considering the strategy of the other teams and justify her decision accordingly (see the third example above). In such cases, this message should be considered as level-2. The acknowledgment of other teamsâ voting strategy may not always be obvious or

may be worded differently such as 'hurting the otherâs decision' or ânot being helpfulâ (see the last three examples above)

Level-3 Player:

Characteristics Assumes that almost all other subjects are level-2 players (partially degenerate beliefs). The reasoning in a level-3 player message will have similarities with a level-1 player message but it will have additional arguments indicating that she assumes others are level-2 players.

Examples 'If everyone else assumes others play their own signal then they will always play red. Since I have the blue ball, it is more likely that we have the blue urn so I will vote blue'

'I think itâs all the same rule. Since I have the blue ball, it is more likely that we have the blue urn so I will vote blue'

Note As stated above, level-3 players are likely to follow their signal like a level-1 player yet they will argue to do so through a much more intricate argument (unlike a level-1 player merely stating probabilities to argue her action). Level-3 players are rare. Higher levels (level-4 etc.) are assumed to not occur; therefore, you should consider only the first 4 levels of thinking.

Thank you.