Origins of Algorithmic Instabilities in Crowdsourced Ranking

KEITH BURGHARDT, USC Information Sciences Institute, USA
TAD HOGG, Institute for Molecular Manufacturing, USA
RAISSA M. D'SOUZA, University of California, Davis, USA and Santa Fe Institute, USA
KRISTINA LERMAN, USC Information Sciences Institute, USA
MÁRTON PÓSFAI, Central European University, Hungary and University of California, Davis, USA

Crowdsourcing systems aggregate decisions of many people to help users quickly identify high-quality options, such as the best answers to questions or interesting news stories. A long-standing issue in crowdsourcing is how option quality and human judgement heuristics interact to affect collective outcomes, such as the perceived popularity of options. We address this limitation by conducting a controlled experiment where subjects choose between two ranked options whose quality can be independently varied. We use this data to construct a model that quantifies how judgement heuristics and option quality combine when deciding between two options. The model reveals popularity-ranking can be unstable: unless the quality difference between the two options is sufficiently high, the higher quality option is not guaranteed to be eventually ranked on top. To rectify this instability, we create an algorithm that accounts for judgement heuristics to infer the best option and rank it first. This algorithm is guaranteed to be optimal if data matches the model. When the data does not match the model, however, simulations show that in practice this algorithm performs better or at least as well as popularity-based and recency-based ranking for any two-choice question. Our work suggests that algorithms relying on inference of mathematical models of user behavior can substantially improve outcomes in crowdsourcing systems.

CCS Concepts: • **Human-centered computing** → **User models**; **Laboratory experiments**; *Heuristic evaluations*; **Empirical studies in visualization**; **Empirical studies in interaction design**.

Additional Key Words and Phrases: decision-formation, crowd-wisdom, algorithmic instability, algorithmic bias

ACM Reference Format:
Keith Burghardt, Tad Hogg, Raissa M. D'Souza, Kristina Lerman, and Márton Pósfai. 2020. Origins of Algorithmic Instabilities in Crowdsourced Ranking. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 166 (October 2020), 20 pages. https://doi.org/10.1145/3415237

1 INTRODUCTION

Crowdsourcing websites aggregate judgments in order to help users discover high quality content. These systems typically combine choices of many people to algorithmically *rank* content so that

Authors' addresses: Keith Burghardt, USC Information Sciences Institute, USA, 4676 Admiralty Way, Marina del Rey, Caliornia, 90292, keithab@isi.edu; Tad Hogg, Institute for Molecular Manufacturing, USA, 555 Bryant Street, Suite 354, Palo Alto, Caliornia, 94301, tadhogg@yahoo.com; Raissa M. D'Souza, University of California, Davis, USA, 1 Shields Ave, Davis, Caliornia, 95616, Santa Fe Institute, USA, 1399 Hyde Park Road, Santa Fe, New Mexico, 87501, raissa@cse.ucdavis.edu; Kristina Lerman, USC Information Sciences Institute, USA, lerman@isi.edu; Márton Pósfai, Central European University, Hungary, Nádor u. 9, Budapest, Hungary, 1051, University of California, Davis, USA, posfaim@ceu.edu.

2573-0142/2020/10-ART166 $15.00
https://doi.org/10.1145/3415237

better items—product reviews [39], news stories [46, 56], or answers on question-answering (Q&A) platforms [5, 65]—are easier to find.

Despite a long history of crowdsourcing [17, 23], some of its limitations have only recently become apparent. Salganik et al. (2006) found that aggregating the votes of many people to rank songs increases the inequality and instability of song popularity. Even when starting from the same initial conditions, the same songs could end up with vastly different rankings. Other studies have shown that algorithmic ranking can amplify the inequality of popularity [35, 37] and bias collective outcomes in crowdsourcing applications [14, 19]. In addition, information about the choices of other users affects decisions in complex ways [29, 46, 58]. Unfortunately for crowdsourcing system designers, it is still not clear how these finding could help improve collective outcomes, in large part due to difficulty of quantifying the quality of options (e.g., the best answer to a question) and its impact on individual decisions.

To better understand and improve collective outcomes that emerge from individual decisions, we break down the crowdsourcing task into its basic elements: item quality, item ranking, and social influence. We create a controlled experiment to study how these elements jointly affect individual decisions and collective outcomes. We use experimental data to construct and validate a mathematical model of human judgements, and then use it to explore algorithmic ranking and identify strategies to improve crowdsourcing performance. Our study addresses the following research questions:

RQ1 How does quality and presentation of options jointly impact individual decisions?
RQ2 When is algorithmic ranking unstable and does not reliably identify the best option?
RQ3 How can we stabilize algorithmic ranking such that the best option is typically ranked first?

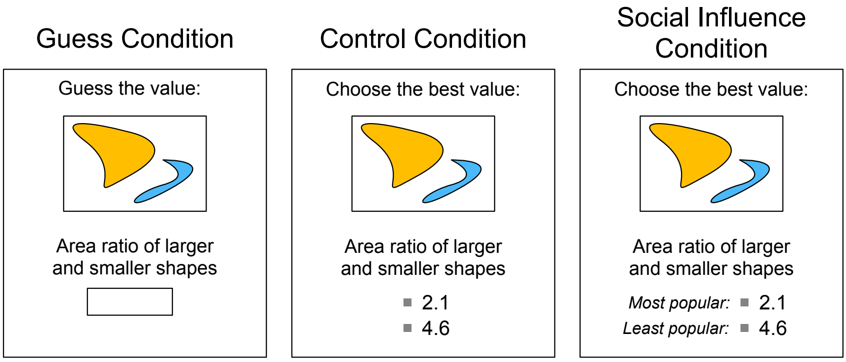

Guess Condition
Guess the value:
Area ratio of larger and smaller shapes
Control Condition
Choose the best value:
Area ratio of larger and smaller shapes
2.1
4.6
Social Influence Condition
Choose the best value:
Area ratio of larger and smaller shapes
Most popular: 2.1
Least popular: 4.6

Fig. 1. Schematic of the experiment conditions. Left: guess condition, where subjects write a guess for the correct value. Center: control condition, where subjects choose the best of two answers. Right: social influence condition, where a randomly chosen answer is called the "most popular" and is ranked first.

Our experiment asks subjects to choose the best answer to questions with objectively correct answers, such as the number of dots in a picture. Figure 1 illustrates one such question, where we ask users to find the area ratio between the largest and smallest shapes. (Other questions used in the experiment are shown in Appendix Fig. 8.) While these simple questions abstract away some of the complexity of the often-subjective decision making people do in crowdsourcing systems, they allow quality to be objectively measured and its effects on decisions better understood.

The experiment has three conditions shown in Fig. 1. In the first condition, we let subjects write their answers to understand how their subjective guesses deviate from the correct answers. In the remaining conditions, we ask subjects to choose the best of two randomly generated answers that are randomly ordered. In the control condition, subjects are not told how they are ordered, while in the social influence condition, the first answer is labeled “more popular”. These simple conditions allow us to disentangle the elements of crowdsourcing systems and begin quantifying how individual decisions affect collective outcomes.

To begin answering RQ1, we construct a mathematical model of the probability to choose an option as a function of its position and quality. The model requires only two parameters to measure *cognitive heuristics* (mental shortcuts people use to make quick and efficient judgements) and is in excellent agreement with experimental data. The first parameter reflects a user’s preference to pick the first answer (known as “position bias” [35–37, 56]) and the other parameter measures the rate at which answers are guessed at random. Subjects otherwise pick an answer closest to their initial (unobserved) guess. We call this model the Biased Initial Guess (BIG) model. The BIG model demonstrates that the “social influence” experiment condition enhances position bias [35], therefore cognitive heuristics that produce position bias and social influence can be quantified with a single parameter, a substantial simplification over previous work [36, 56]. Moreover, it helps explain why users often choose the worst answer when answer quality differences are small.

The BIG model not only improves our understanding of how answers are chosen, but it allows us to test different ranking policies in simulations to answer RQ2 and RQ3. Importantly, these simulations demonstrate that cognitive heuristics can make popularity-based ranking highly unstable. When the quality difference between the options is small, initially minor differences in popularity can create a cumulative advantage [61], meaning the better answer does not always become the most popular. However, when the quality difference passes a critical point, popularity-ranking is stable, and the better answer eventually becomes the most popular. These results may help explain an under-appreciated finding of Salganik et al. (2006) that the best and worst songs tended to be correctly ranked when songs were ordered by popularity, but intermediate-quality songs landed anywhere in between.

Finally, to answer RQ3, we propose an algorithm that rectifies this instability by ordering answers based on their inferred quality, which we call RAICR: Rectifying Algorithmic Instabilities in Crowdsourced Ranking. This method is found to be stable and consistently order the better answer first, thus making best answers easier to find, even when they are only slightly better. RAICR ranks answers as well as, or better than, common baselines such as ordering by popularity or by recency (ranking by the last answer picked).

Our work shows that individual decisions within crowdsourcing systems are strongly affected by cognitive heuristics, which collectively create instability and poor crowd wisdom. Designers of crowdsource systems need to account for these biases in order make good content easier to find. Algorithms such as RAICR, however, can correct for these biases, thereby improving the wisdom of crowds.

2 RELATED LITERATURE

2.1 Crowdsourcing

Crowdsourcing has a two-century long history demonstrating how a collective can outperform individual experts [17, 23, 33, 55, 57], thus creating the moniker “wisdom of crowds”. Crowds have been shown to beat sport markets [13, 50], corporate earnings forecasts [16], and improve visual searches [32]. One reason crowd wisdom works is due to the law of large numbers: assuming unbiased and independent guesses, the average guess should converge to the true value.

Individual decisions, including in online settings, are biased by cognitive heuristics, such as anchoring [54], primacy [42], prior beliefs [63] and position bias [35]. These biases are not necessarily canceled out with large samples [34, 51]. As a result, aggregating guesses of a crowd does not necessarily converge to the correct answer.

Guesses are usually not independent, which can sometimes improve the wisdom of crowds. Social influence models, such as the DeGroot model [18], have been shown to push simulated agents to an optimal decision [4, 8, 25, 45]. These results have been backed up experimentally [9, 10, 59, 60], even when opinion polarization is included [10]. One reason social influence can be beneficial is that it encourages people who are way off the mark to improve their guess [3, 9, 44].

Often, however, social influence can reduce crowd wisdom. Corporate earnings predictions [16], jury decisions [15, 33], and other guesses can degrade with influence [40, 41, 55], and malevolent individuals can manipulate people to make particular collective decisions [6, 46]. Too much influence by a single individual can also reduce the wisdom of collective decisions [4, 9, 25], and deferring to friends can sometimes make unpopular (and potentially low-quality) ideas appear popular [38].

Recent work has also demonstrated how other cognitive heuristics can affect crowd wisdom. After the landmark study by Salganik et al. (2006), some researchers found that social influence has no effect on decisions [43], or that position bias, i.e., the preference to choose options listed first, largely explain biases in crowdsourcing [36, 37]. Social influence instead enhances the position bias [35, 36]. Burghardt et al. (2018) have begun to tease apart these effects, showing that while social influence enhances position bias, it has no marginal effect when we control for position. This is consistent with our approach, which models and rectifies both biases with a single parameter.

2.2 Algorithmic Ranking

The goal of a ranking algorithm is to make good content easier to find. Many papers have begun to address this goal [11, 31, 49], which has recently been applied to crowdsourced ranking. Because of human biases, however, algorithms that naïvely use human feedback to suggest content will end up forming echo chambers [12, 26, 28], or only recommend already-popular items [1, 2]. This can also give some content a cumulative advantage [61], even when it is of similar quality to content that remains unpopular.

To correct for algorithmic bias in this paper, we create a ranking method that follows the strategy of Watts, who says, "...we can instead measure directly how they respond to a whole range of possibilities and react accordingly" [62]. In the present context, this strategy implies we can create better algorithms for option ranking by observing, and addressing, how people respond to social influence and position biases. We show this strategy applied in RAICR improves upon simple algorithms used in the past, which include ordering results by popularity [35] or recency [37]. While some crowdsourced ranking strategies use a two-tier platform model, in which researchers rank options based on whether content is downloaded and rated [3, 43, 52], RAICR is based on a common simpler model in which we only observe if content is chosen [35, 56]. This subtle difference implies many previous ranking schemes are not applicable. The present paper also compliments previous work that uses features, such as answer position, to predict Q&A website quality [14, 53].

3 EXPERIMENT

The experiment asked subjects, hired through Amazon Mechanical Turk between August 2018 and September 2019, to answer a series of questions shown in the Appendix. The order of questions was randomized for each subject, and questions were not time-limited. Questions include specifying the ratio of the areas of two shapes or the lengths of two lines, or the number of dots in an image. We designed the experiment around quotidian tasks that do not require specific expertise but are difficult for most people. Despite their difficulty, the questions have objectively correct answers.

Table 1. Number of subjects for each guess question (after cleaning)

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
595	594	593	595	592	593	589	591	584	590

We quantify the quality of an answer by the difference from the mean of all guesses, which better matches a log-normal distribution, as discussed later. In the Appendix, we define quality of an answer by the difference from the correct value and find results are qualitatively very similar.

Subjects were assigned to one of three conditions. In the *guess condition*, shown in the left panel of Fig. 1, the subjects could freely type their guesses. In the *control condition*, shown in the center panel of Fig. 1, subjects were told to choose the best among two options, and were not told how the two answers were ordered. Finally, in the *social influence condition*, shown in the right panel of Fig. 1, they were told the first among two answers was the most popular, but the layout was otherwise identical to the previous condition. We only showed ten questions to each subject to reduce the effects of performance depletion in Q&A systems [20]. Further, to reduce bias due to other phenomena, such as the decoy effect [30], we showed subjects only two choices. Subjects in the same condition but assigned on different days have statistically similar behavior.

Approximately 1800 subjects are evenly split between the conditions (596, 586, and 587 for the guess, control, and social influence condition, respectively). For guess values, we remove extreme outliers (guesses smaller than 1 or greater than 10^6). All answers are supposed to be greater than 1, while values greater than 10^6 may affect mean values and appear to represent throwaway answers. The number of valid participants for each question is shown in Table1.

Mechanical Turk workers were hired if they had an approval rate of over 95%, completed more than 1000 Human Intelligence Tasks (HITs), and never participated in any of the experiment conditions before. Each worker was paid $1.00 for the guessing condition and $0.50 for the other two conditions. The assignment took 6 minutes on average for the guess condition and 2 minutes on average for the other conditions, equivalent to an hourly wage of $10 − $15. The human experiment was approved by the appropriate IRB board.

4 RESULTS

4.1 A Mathematical Model of Decisions

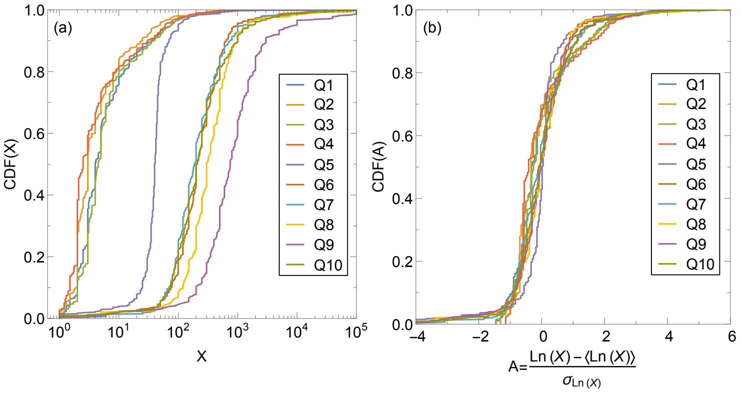


Fig. 2. The CDF of guesses for each question. (a) Before normalization, and (b) after normalization.

We use data gathered from the experiment to answer **RQ1**: *How does quality and presentation of options jointly impact individual decisions?*

4.1.1 Open-Ended Experiment Condition. To derive the model of how people make these decisions, we start by constructing the distribution of guesses for each question (the *guess* condition in the experiment). The guesses, plotted in Appendix Fig. 9 and CDF shown in Fig. 2a, are highly variable (by as many as six orders of magnitude), while the correct answers vary by three orders of magnitude. The median guess may differ from the true answer, in agreement with previous work [34], but values are typically the correct order of magnitude.

We normalize guesses by defining a new variable A:

$$A = \frac{\ln(X) - \langle \ln(X) \rangle}{\sigma_{\ln(X)}},$$

where X is a guess value, and $\langle \ln(X) \rangle$ is the mean of the logarithm of guesses. Figure 2b shows that this simple normalization scheme effectively collapses answer guesses to a single distribution. We show in the Appendix that guesses are not normally distributed, but instead are better approximated as log-normal, in agreement with previous work on a different set of questions [34]. The normalized guesses A can be thought of as the z-scores in log-normal distributions. Alternative ways to center data, shown in the Appendix, produce similar results. We use these normalized guesses and distributions in the remaining two experiment conditions. Intuitively, if the mean of all guesses converges to the correct answer, $A = 0$ can be thought of as the best answer.

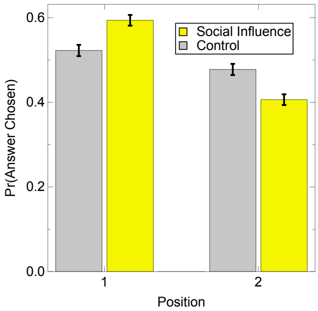


Fig. 3. Probability to choose an answer versus its position for the social influence and control condition.

4.1.2 Two-Choice Experiment Conditions. The latter two experiment conditions require subjects to pick the best among two answers to the question. For both the control and social influence conditions, answers are ordered vertically, with one answer above the other. There is a significant position effect when answers are ordered this way, as shown in Fig. 3. In the control condition, there is a slightly greater probability (52%) to choose the first (top) answer over the last one (p-value < 0.001). In the social influence condition, meanwhile, the probability to choose the first answer is substantially larger (59%) and statistically significantly different than the control condition (p-value < 0.001). This is in agreement with previous work showing that social influence amplifies the position effect [35].

4.1.3 The Biased Initial Guess Model. We now have the necessary ingredients to model how decisions to choose an answer are affected by its quality, position, and social influence. We present the Biased Initial Guess (BIG) decision model and show it is consistent with the data.

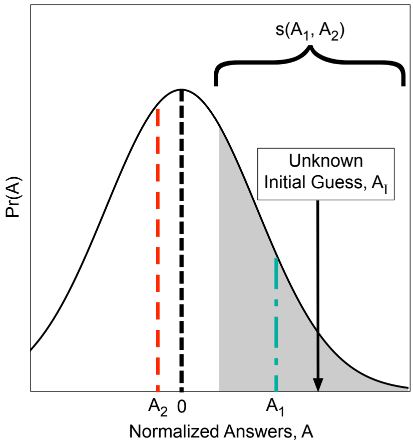


Fig. 4. Schematic to calculate $s(A_1, A_2)$. Guesses follow an approximately log-normal distribution unique to each question, but the normalized guesses, A are approximately standard normal distributed. In an experiment, assume two candidate answers are provided, A_1 is listed first, and A_2 is listed last. The variable $s(A_1, A_2)$ in Eq. 2 is the probability an unknown initial guess is closer to A_1 than A_2.

We first discuss the simplest case where a user has to choose the better of two answers A_1, listed first, and A_2, listed last, in the absence of cognitive biases. Figure 4 shows probability of the answer, with the normalized values of the choices A_1 and A_2, as well as the user's initial guess A_I about the true answer, which we do not observe. All things equal, the user will choose the first answer A_1 if it is closer to the initial guess, i.e., if $|A_1 - A_I| < |A_2 - A_I|$, and will otherwise choose A_2. The probability to choose A_1 is then:

$$s(A_1, A_2) = \begin{cases} \Pr(A_I > \frac{A_1+A_2}{2}) & A_1 > A_2 \\ 1/2 & A_1 = A_2 \\ \Pr(A_I < \frac{A_1+A_2}{2}) & A_1 < A_2 \end{cases} \tag{1}$$

Assuming A_1 and A_2 follow a normal distribution quantified in the guess experiment condition,

$$s(A_1, A_2) = \begin{cases} \operatorname{erfc}\left(\frac{1}{\sqrt{2}} \frac{A_1+A_2}{2}\right)/2 & A_1 \geq A_2 \\ \left[1 + \operatorname{erf}\left(\frac{1}{\sqrt{2}} \frac{A_1+A_2}{2}\right)\right]/2, & A_1 < A_2 \end{cases} \tag{2}$$

where erf(.) is the error function, erfc(.) = 1 − erf(.) is the complimentary error function. When $s(A_1, A_2) > 0.5$, A_1 is not only more likely to be closer to the initial guess (A_I) than A_2, but is also closer to zero than A_2, and therefore the objectively better answer. On the other hand, when $s(A_1, A_2) < 0.5$, A_1 is further from most guesses compared to A_2 and the objectively worse answer.

To better model decision-making, we have to account for biases, due to cognitive heuristics and algorithmic ranking, to explain why people do not always choose the best option. As shown in Fig. 3, sometimes they choose the first answer even if it is not the best answer. We quantify this position bias by assuming that with probability p participants choose the first answer regardless of its quality. This parameter should presumably be small in the control condition and large in the social influence condition. Subjects may also choose an answer regardless of its position or quality because there is no monetary incentive to choose good answers. We model this by allowing subjects to choose an answer at random with a probability r. Taking these two heuristics into account, we arrive at the BIG Model:

$$\begin{aligned}\Pr(\text{Choose } A_1) &= r/2 + (1-r)[p + (1-p)s(A_1, A_2)]\\ &= \begin{cases} r/2 + (1-r)\left[p + (1-p)\mathrm{erfc}\left(\frac{1}{\sqrt{2}}\frac{A_1+A_2}{2}\right)/2\right] & A_1 \geq A_2\\ r/2 + (1-r)\left[p + (1-p)\left[1 + \mathrm{erf}\left(\frac{1}{\sqrt{2}}\frac{A_1+A_2}{2}\right)\right]/2\right], & A_1 < A_2 \end{cases}\end{aligned} \qquad (3)$$

The probability of choosing A_2 is simply the compliment of this probability. Because $A = 0$ is expected to be the best answer, with some simple manipulation we can infer the probability the best answer is chosen.

$$\Pr(\text{Choose } A_1|A_1 \text{ Best}) = \begin{cases} r/2 + (1-r)\left[p + (1-p)\mathrm{erfc}\left(\frac{1}{\sqrt{2}}\frac{A_1+A_2}{2}\right)/2\right] & \frac{A_1+A_2}{2} < 0\\ r/2 + (1-r)\left[p + (1-p)\left(1 - \mathrm{erfc}\left(\frac{1}{\sqrt{2}}\frac{A_1+A_2}{2}\right)/2\right)\right] & \frac{A_1+A_2}{2} \geq 0 \end{cases} \qquad (4)$$

A similar equation can model Pr(Choose $A_2|A_2$ Best).

Agreement between the model and data is shown in Fig. 5. In the *control condition* (Fig. 5a), the best parameters are $r = 0.28 \pm 0.02$ and $p = 0.05 \pm 0.02$. We find that the log-likelihood of the model, $\ell = -3002.49$, is not statistically different from log-likelihood if the data came from the model itself: p-value = 0.40. See Methods for how p-values and error bars are calculated. We also check if we need both parameters, r and p, using the likelihood ratio test and Wilks' Theorem [64]. We compare the likelihood ratio of the two-parameter model to simpler models with p or r (or both) set to zero. The probability a simpler model could fit the data as well or better is ≤ 0.002. We conclude that our model describes the control condition very well.

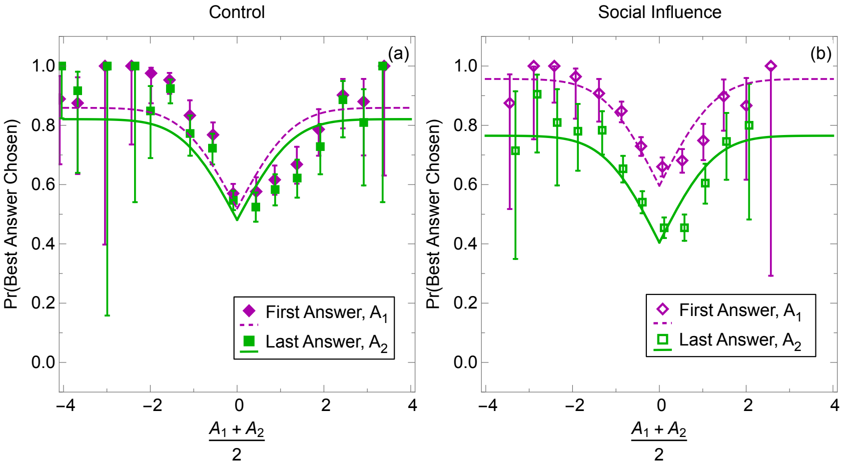


Fig. 5. Decision model agreement with experiment data. Plots show the probability the best answer is chosen for the model (lines) and experiment (symbols) under (a) the control condition and (b) the social influence condition. Purple diamonds are probabilities when the best answer is the first answer, A_1 and green squares for when the best answer is A_2.

The agreement between data and model is similarly close in the *social influence condition* (Fig. 5b). The position bias parameter $p = 0.21 \pm 0.01$ is larger than in the control condition, in agreement with expectations. We also find $r = 0.08 \pm 0.02$, thus social influence reduces the frequency of random guesses. In both experiment conditions, surprisingly, $\approx$ 20% of users choose answers for reasons besides "quality" ($r/2 + (1 - r)p = 18\%$ and 23% for the control and social influence conditions, respectively). Similar to the control condition, we find that the model is consistent with the data. The log-likelihood of the empirical data ($\ell = -3190.13$) is not statistically different from log-likelihood values if the data came from the model: p-value 0.47. The probability a simpler model (r or p set to zero) could fit the data as well or better is $< 10^{-5}$. In conclusion, we find the BIG model is consistent with both experiment conditions and its parameters are interpretable and meaningful. In the Appendix, we show that all these results are consistent when we look at a subset of experiment questions or center the data differently.

4.2 Algorithmic Ranking Instability

Crowdsourcing websites automatically highlight what they consider the best choices to help their users more quickly discover them. For example, Stack Exchange (like other Q&A platforms) usually ranks answers to questions by the number of votes they receive. Despite problems with popularity-based ranking identified in previous studies [37, 52], it is widely used for ranking content in crowdsourcing websites. In this section we identify an instability in popularity-based ranking: the first few votes a worse answer receives can lock it in the top position, where it acquires cumulative advantage [61]. This allows us to answer *RQ2: When is algorithmic ranking unstable and does not reliably identify the best option?*

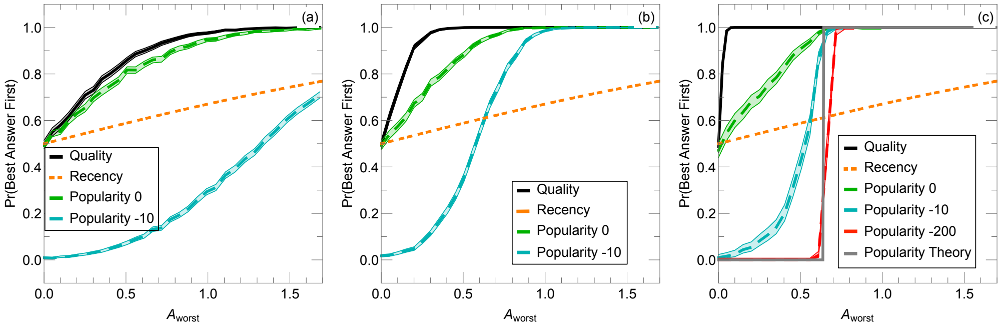


Fig. 6. Comparison of ranking policies via simulations. Plots show the probability the best answer is ranked first after (a) 50, (b) 500, and (c) 20,000 votes, when answers are ranked by quality (black line), recency (orange dashed line), and popularity. Also shown in (c) is the critical value of A_{worst} based on Eqs. 8 and 2. In these simulations, subjects choose answers following the BIG Model with $p = 0.2$ and $r = 0.09$. "Popularity 0" (green dashed line), "Popularity -10" (cyan dashed line), and "Popularity -200" (red dashed line) means that the worst answer starts with a 0, 10, or 200 vote advantage, respectively. Shaded areas are 95% confidence intervals.

To demonstrate the instability, we simulate a group of agents who choose answers according to the BIG model with $p = 0.2$ and $r = 0.09$. In the simulations, one answer is objectively best, e.g., exactly equal to the correct answer ($A_{\text{best}} = 0$), while the worst answer is larger than A_{best} (results are symmetric if $A_{\text{worst}} < A_{\text{best}}$). At each timestep, a new user arrives and independently chooses the first or last answer according to Eq. 3. Answers reorder depending on the ranking algorithm. Figure 6 shows our results. If the worst answer, A_{worst}, is not too large and has a few extra votes (a common occurrence when the worst answer is posted before the better answer), we find that popularity ordering completely breaks down—the worst answer usually becomes more popular and is ranked first. Even when both answers start with the same number of votes, the worse answer is often ranked first. Results remain stable even after 20K votes; the effect does not appear transient.

We can explain this result using our model. Using the compliment of Eq. 3, we can define the probability the best answer is chosen, conditional on it being ranked last:

$$\Pr(\text{Choose } A_{\text{best}} = A_2) = r/2 + (1-r)(1-p)s(A_1, A_2) \tag{5}$$

We find that

$$\Pr(\text{Choose } A_{\text{best}} = A_2) = \Pr(\text{Choose } A_{\text{best}} = A_1) - p(1-r) \tag{6}$$

where $\Pr(\text{Choose } A_{\text{best}} = A_1)$ is the probability of choosing the best answer when it is ranked first. If

$$\Pr(\text{Choose } A_{\text{best}} = A_2) > \Pr(\text{Choose } A_{\text{worst}} = A_1) \tag{7}$$

then subjects are more likely to choose the better answer regardless of whether its ranked first or second, therefore answer order is stable. When the above inequality is not true, however, then subjects are more likely to pick the first answer regardless of its quality, and the popularity ranking is unstable. The critical value of A_{worst} between the two regimes is when

$$s_{\text{crit}} = \frac{1}{2(1-p)} \tag{8}$$

and is independent of r. Intuitively, if the answer is exceptionally bad, it will always be less popular. This is alike to the results in Salganik's MusicLab study [52], where particularly good and bad songs were ranked correctly. However, if the first answer is likely to be chosen regardless of quality, the worst answer will continue accumulating votes and remain in the top position. Given $A_{\text{best}} = 0$, we can use Eqs. 8 and 2 to numerically solve for the critical value of A_{worst}. We plot the critical point as a function of p and A_{worst} in phase diagram in Fig. 7. We see that there is always a large part of the phase space where popularity-based ranking will be unpredictable if answer quality is close together. Based on Eq. 8, if $p \geq 0.5$, there will be no case where popularity-based ranking is guaranteed to correctly rank answers. While this is an extreme case, it still points to substantial limitations of popularity-based ranking.

4.3 Stabilizing Algorithmic Ranking

Given the problems with popularity ranking, it is critical to create a more consistent ranking algorithm. Moreover, while we have so-far explored questions with numeric answers, we want a method that works for all types of answers. Using the BIG model, we can answer RQ3: *How can we stabilize algorithmic ranking such that the best option is typically ranked first?*

Assume we can approximate p and r, then we can invert Eq. 3 and use votes to infer the only unknown variable $s(A_{\text{best}}, A_{\text{worst}})$. When $s(A_{\text{best}}, A_{\text{worst}}) > 0.5$, A_{best} is the best answer, but if we incorrectly rank A_{worst} first, $s(A_{\text{worst}}, A_{\text{best}}) < 0.5$. We can therefore rank the answer in which $s(A_{\text{best}}, A_{\text{worst}}) > 0.5$ as the best answer. This is the backbone of the RAICR algorithm. The algorithm uses maximum likelihood estimation to solve for $s(A_{\text{best}}, A_{\text{worst}})$, as shown in the Appendix. Therefore, if the data matches the BIG model with the correct r and p parameters, our method *optimally infers the correct raking* by having minimal variance and no bias [48]. Moreover, *this method only depends on the votes an answer receives rather than the type of answer*, such as a numerical or textual answer. We compare quality ranking to popularity-based ranking, and recency-based ranking (ranking by the last answer picked), as shown in Fig. 6. The probability recency ranks the best answer first is calculated as the self-consistent equation:

$$\begin{aligned}\Pr(\text{Rank } A_{\text{best}} \text{ First}|\text{Recency}) &= \Pr(\text{Choose } A_{\text{best}}|A_{\text{best}} \text{ First})\Pr(\text{Rank } A_{\text{best}} \text{ First}|\text{Recency}) \\ &+ \Pr(\text{Choose } A_{\text{best}}|A_{\text{best}} \text{ Last})(1 - \Pr(\text{Rank } A_{\text{best}} \text{ First}|\text{Recency})).\end{aligned} \tag{9}$$

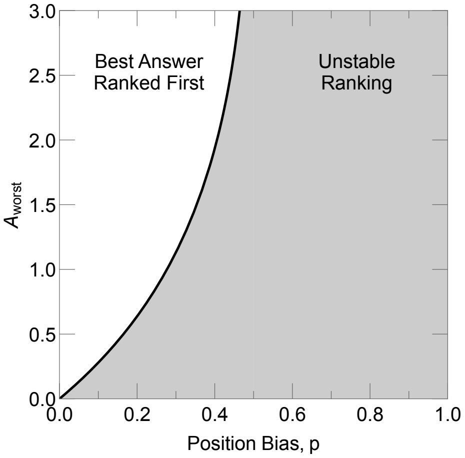


Fig. 7. Phase diagram for popularity-based ranking based on Eqs. 8 and 2. We show the boundary between regimes where best answers are guaranteed to be ranked first with enough votes (white area) and where ranking is unstable (gray area) as a function of position bias (p) and the value of the worse answer (A_{worst}). In this plot, larger values of $A_{\text{worst}} > 0$ correspond to worse answers and results are symmetric for $A_{\text{worst}} < 0$.

This represents the limit answers acquire many votes. The solution to this equation is:

$$\Pr(\text{Rank } A_{\text{best}} \text{ First}|\text{Recency}) = \frac{2(1-p)(1-r)s(A_{\text{best}}, A_{\text{worst}}) + r}{2 - 2p(1-r)} \tag{10}$$

We find that the RAICR algorithm performs at least as well as popularity-based ranking by ranking the better answer first, and often better after 20-50 votes (Fig. 6a and Appendix Fig. 12). Moreover, RAICR always outperforms the recency-based algorithm. The benefit of the RAICR algorithm only improves as we collect more votes. For example, Fig. 6b shows after after 500 votes the advantage of RAICR is larger, and Fig. 6c shows that after 20K votes the method is nearly optimal. Popularity-based ranking performs badly for $A_{worse} < 0.6$, and recency performs worst if $A_{worse} > 0.6$.

While we show these results hold when we have 20 to 20,0000 votes, many platforms are underprovisioned, with a large fraction of webpages receiving little attention and votes [7, 24]. It is the webpages that receive many votes, however, which may be the most important. Correctly ranking options in these popular pages is therefore especially critical to crowdsourcing websites. Moreover, a moderate number of votes is generally needed to make a reasonable estimate of quality, so very few methods will accurately rank unpopular pages.

One caveat of the RAICR algorithm is that we need an approximate value for two parameters: r and p. What happens if either of these parameters are far off? For example, we could assume $r = 0$ when $r = 0.09$. Results in the Appendix, however, show that the findings are quantitatively very similar, and therefore our model is robust to assumptions about r. On the other hand, what if we incorrectly estimate both r and p? We show in the Appendix that even in this worst-case scenario, our method performs slightly worse, but is comparable or substantially better than popularity-based ranking.

5 FUTURE WORK AND DESIGN IMPLICATIONS

In the experiment, we decomposed the crowdsourcing task to its basic components to reduce the complexity and variation inherent in real world tasks. While this helps to disentangle effects of option quality and position without confounding factors muddying the relationships, future work is needed to verify ecological validity of our results [27]. It is encouraging that the probability to choose an answer (Fig. 3b) is quantitatively similar to empirical data gathered from Stack Exchange [35], despite Mechanical Turk workers not being representative of the general population [47].

For simplicity, we only explored two-option questions; future work should aim to understand multi-option decision-making given options of variable quality. A generalization of RAICR should also address more complicated biases, such as preference for round numbers [21], anchoring [22, 54], and biases that appear in multi-option decisions, such as the decoy effect [30]. Finally, while RAICR is found to be robust to moderate changes in its parameters, this algorithm and its extensions may fail to rank options properly if its parameters are far off, or if the BIG model is wrong. In the experiment, the model is backed by data, but future work needs to address whether other tasks or questions follow this model.

Our results offer implications for crowdsourcing platforms. First, designers must recognize the limits of crowdsourcing due to biases implicit in their platform. In our experiment people often upvoted options at random (up to 20% of all votes), and chose an inferior option simply because it was shown first. This creates a ranking instability when options are of similar quality. Our controlled experiment and mathematical model point to ways we can counteract this instability. Designers should similarly create platform-tailored mathematical models and controlled experiments to rigorously test how crowds can better infer the best options.

A key property of our RAICR algorithm is that it relies on accurate modeling of user decisions to counteract cognitive biases. In effect, each vote is weighted depending on the ranks of answers at the time the vote is cast. The idea is similar to one described by Abeliuk et al. 2017 that ranks items by their inferred quality in order to more robustly identify blockbuster items. Similar weighting schemes could be applied to future debiased algorithms to address the unique goals of each crowdsourcing platform.

There are also simple methods that platforms like Reddit, Facebook, and Stack Exchange can try that may greatly outperform the baselines we mention in our paper. For example, items that have not yet acquired many votes can be ranked randomly to reduce initial ranking biases. Alternatively, new posts and links could be ranked appropriately but their popularity could be hidden until they gather enough votes. Our results suggest this could reduce social influence-based position bias up until the true option quality is more obvious.

6 CONCLUSION

In this paper, we introduce an experiment designed to inform how cognitive biases and option quality interact to affect crowdsourced ranking. Results from this experiment help us create a novel mathematical decision model, the BIG model, that greatly improves our understanding of how people find the best answer to a question as a function of answer quality, rank, and social influence. This model is then applied to the RAICR algorithm to better rank answers. The BIG model also helped us uncover instability in popularity-based ranking. The instability depends on the quality of options: when there are large differences between option qualities, popularity converges optimally and predictably. However, when the difference between the quality of options is small, the better option may not always become the most popular. These results can help us better understand the foundational empirical results of Salganik et al. (2006), who found that popularity-based ranking correctly ranked high and low quality songs, while the ranking of intermediate quality songs was highly unstable. Although our experimental setup is undeniably simpler than real crowdsourcing websites, our results suggest that accurate models of user behavior together with mathematically principled inference can improve the efficiency of crowdsourcing.

ACKNOWLEDGMENTS

Our work is supported by the US Army Research Office MURI Award No. W911NF-13-1-0340 and the DARPA Award No. W911NF-17-1-0077. Data as well as code to create experiments, create simulations, and analyze data is available at https://github.com/KeithBurghardt/QualityRankCodeAndData.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 42–46. https://doi.org/10.1145/3109859.3109912

[2] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing Popularity Bias in Recommender Systems with Personalized Re-ranking. *arXiv preprint: 1901.07555* (2019).

[3] Andrés Abeliuk, Gerardo Berbeglia, Pascal Van Hentenryck, Tad Hogg, and Kristina Lerman. 2017. Taming the Unpredictability of Cultural Markets with Social Influence. In *Proceedings of the 26th International World Wide Web Conference (WWW2017)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland.

[4] Daron Acemoglu, Munther A. Dahleh, Ilan Lobel, and Asuman Ozdaglar. 2011. Bayesian Learning in Social Networks. *The Review of Economic Studies* 78, 4 (03 2011), 1201–1236. https://doi.org/10.1093/restud/rdr004 arXiv:http://oup.prod.sis.lan/restud/article-pdf/78/4/1201/18376066/rdr004.pdf

[5] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proceedings of the 17th international conference on World Wide Web*. ACM, New York, NY, 665–674.

[6] S. E. Asch. 1951. *Effects of group pressure upon the modification and distortion of judgments.* Carnegie Press, Oxford, UK. 177–190 pages.

[7] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[8] Venkatesh Bala and Sanjeev Goyal. 1998. Learning from Neighbours. *The Review of Economic Studies* 65, 3 (1998), 595–621. http://www.jstor.org/stable/2566940

[9] Joshua Becker, Devon Brackbill, and Damon Centola. 2017. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences* 114, 26 (2017), E5070–E5076. https://doi.org/10.1073/pnas.1615978114 arXiv:https://www.pnas.org/content/114/26/E5070.full.pdf

[10] Joshua Becker, Ethan Porter, and Damon Centola. 2019. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences* 116, 22 (2019), 10717–10722. https://doi.org/10.1073/pnas.1817195116

[11] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. 2011. Quality-Biased Ranking of Web Documents. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (Hong Kong, China) *(WSDM '11)*. Association for Computing Machinery, New York, NY, USA, 95–104. https://doi.org/10.1145/1935826.1935849

[12] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15, 3 (01 Sep 2013), 209–227. https://doi.org/10.1007/s10676-013-9321-6

[13] Alasdair Brown and J. James Reade. 2019. The wisdom of amateur crowds: Evidence from an online community of sports tipsters. *European Journal of Operational Research* 272, 3 (2019), 1073 – 1081. https://doi.org/10.1016/j.ejor.2018.07.015

[14] Keith Burghardt, Emanuel F. Alsina, Michelle Girvan, William Rand, and Kristina Lerman. 2017. The Myopia of Crowds: A Study of Collective Evaluation on Stack Exchange. *PLOS ONE* 12, 3 (2017), e0173610.

[15] Keith Burghardt, William Rand, and Michelle Girvan. 2019. Inferring models of opinion dynamics from aggregated jury data. *PLoS ONE* 14, 7 (2019), e0218312.

[16] Zhi Da and Xing Huang. 2019. Harnessing the Wisdom of Crowds. *Management Science* 0, 0 (2019), 1–21. https://doi.org/10.1287/mnsc.2019.3294

[17] Marquis de Condorcet. 1976. *"Essay on the Application of Mathematics to the Theory of Decision-Making." Reprinted in Condorcet: Selected Writings.* Bobbs-Merrill,, Indianapolis, Indiana.

[18] Morris H. Degroot. 1974. Reaching a Consensus. *J. Amer. Statist. Assoc.* 69, 345 (1974), 118–121. https://doi.org/10.1080/01621459.1974.10480137

[19] Himel Dev, Karrie Karahalios, and Hari Sundaram. 2019. Quantifying Voter Biases in Online Platforms: An Instrumental Variable Approach. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 120.

[20] Emilio Ferrara, Nazanin Alipoufard, Keith Burghardt, Chiranth Gopal, and Kristina Lerman. 2017. Dynamics of Content Quality in Collaborative Knowledge Production. In *ICWSM '17 Proceedings of the 11th International AAAI Conference on Web and Social Media.*

[21] Catherine Fitzmaurice and Ken Pease. 1986. *The psychology of judicial sentencing.* Manchester University Press, Manchester, UK.

[22] Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics* 40, 1 (2011), 35 – 42. https://doi.org/10.1016/j.socec.2010.10.008

[23] F. Galton. 1908. Vox Populi. *Nature* 75 (1908), 450–451.

[24] Eric Gilbert. 2013. Widespread Underprovision on Reddit. In *CSCW '13: Proceedings of the 2013 conference on Computer supported cooperative work*. Association for Computing Machinery, 803–808.

[25] Benjamin Golub and Matthew O. Jackson. 2010. Naïve Learning in Social Networks and the Wisdom of Crowds. *American Economic Journal: Microeconomics* 2, 1 (2010), 112–49.

[26] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 2125–2126. https://doi.org/10.1145/2939672.2945386

[27] Daniel Herbst and Alexandre Mas. 2015. Peer effects on worker output in the laboratory generalize to the field. *Science* 350, 6260 (October 2015), 545–549.

[28] Martin Hilbert, Saifuddin Ahmed, Jaeho Cho, Billy Liu, and Jonathan Luu. 2018. Communicating with Algorithms: A Transfer Entropy Analysis of Emotions-based Escapes from Online Echo Chambers. *Communication Methods and Measures* 12, 4 (2018), 260–275. https://doi.org/10.1080/19312458.2018.1479843

[29] T. Hogg and K. Lerman. 2015. Disentangling the effects of social signals. *Human Computation Journal* 2, 2 (2015), 189–208.

[30] Joel Huber, John W. Payne, and Christopher Puto. 1982. Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *Journal of Consumer Research* 9, 1 (06 1982), 90–98. https://doi.org/10.1086/208899 arXiv:http://oup.prod.sis.lan/jcr/article-pdf/9/1/90/5205641/9-1-90.pdf

[31] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418

[32] Mordechai Z. Juni and Miguel P. Eckstein. 2017. The wisdom of crowds for visual search. *Proceedings of the National Academy of Sciences* 114, 21 (2017), E4306–E4315. https://doi.org/10.1073/pnas.1610732114

[33] Serguei Kaniovski and Alexander Zaigraev. 2011. Optimal jury design for homogeneous juries with correlated votes. *Theory Dec.* 71 (2011), 439–459.

[34] Albert B. Kao, Andrew M. Berdahl, Andrew T. Hartnett, Matthew J. Lutz, Joseph B. Bak-Coleman, Christos C. Ioannou, Xingli Giam, and Iain D. Couzin. 2018. Counteracting estimation bias and social influence to improve the wisdom of crowds. *Journal of The Royal Society Interface* 15, 141 (2018), 20180130. https://doi.org/10.1098/rsif.2018.0130

[35] Tad Hogg Keith Burghardt and Kristina Lerman. 2018. Quantifying the Impact of Cognitive Biases in Crowdsourcing. In *Proceedings of The 12th International AAAI Conference on Web and Social Media (ICWSM-18)*. AAAI.

[36] Coco Krumme, Manuel Cebrian, Galen Pickard, and Sandy Pentland. 2012. Quantifying Social Influence in an Online Cultural Market. *PLoS ONE* 7, 5 (2012), e33785.

[37] K. Lerman and T. Hogg. 2014. Leveraging position bias to improve peer recommendation. *PLOS ONE* 9, 6 (2014), e98914.

[38] K Lerman and X-Z Yan, X amd Wu. 2016. The "Majority Illusion" in Social Networks. , e0147617 pages.

[39] Young-shin Lim and Brandon Van Der Heide. 2015. Evaluating the wisdom of strangers: The perceived credibility of online consumer reviews on Yelp. *Journal of Computer-Mediated Communication* 20, 1 (2015), 67–82.

[40] Jan Lorenz, Heiko Rauhut, and Bernhard Kittel. 2015. Majoritarian democracy undermines truth-finding in deliberative committees. *Research & Politics* 2, 2 (2015), 2053168015582287. https://doi.org/10.1177/2053168015582287

[41] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* 108, 22 (2011), 9020–9025.

[42] A. Mantonakis, P. Rodero, I. Lesschaeve, and R. Hastie. 2009. Order in Choice: Effects of Serial Position on Preferences. *Psychol. Sci.* 20, 11 (2009), 1309–1312.

[43] Erisa Terolli Marzia Antenore, Alessandro Panconesi. 2018. Songs of a Future Past — An Experimental Study of Online Persuaders. In *Twelfth International AAAI Conference on Web and Social Media*. AAAI.

[44] Pavlin Mavrodiev, Claudio J. Tessone, and Frank Schweitzer. 2013. Quantifying the effects of social influence. *Scientific Reports* 3, 1 (2013), 1360. https://doi.org/10.1038/srep01360

[45] Elchanan Mossel, Allan Sly, and Omer Tamuz. 2015. Strategic Learning and the Topology of Social Networks. *Econometrica* 83, 5 (2015), 1755–1794. https://doi.org/10.3982/ECTA12058

[46] Lev Muchnik, Sinan Aral, and Sean J. Taylor. 2013. Social Influence Bias: A Randomized Experiment. *Science* 341 (2013), 647–651.

[47] K Munger, M Luca, J Nagler, and J Tucker. 2019. Age matters: Sampling strategies for studying digital media effects. (2019). https://osf.io/sq5ub/

[48] Whitney K. Newey and Daniel McFadden. 1994. *Chapter 36: Large sample estimation and hypothesis testing*. Vol. 4. Elsevier Science. 2111–2245 pages.

[49] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/ Previous number = SIDL-WP-1999-0120.

[50] Thomas Peeters. 2018. Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting* 34, 1 (2018), 17 – 29. https://doi.org/10.1016/j.ijforecast.2017.08.002

[51] Dražen Prelec, H. Sebastian Seung, and John McCoy. 2017. A solution to the single-question crowd wisdom problem. *Nature* 541, 7638 (2017), 532–535. https://doi.org/10.1038/nature21054

[52] M. Salganik, P. Dodds, and D. Watts. 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* 311 (2006), 854–856.

[53] Chirag Shah and Jefferey Pomerantz. 2010. Evaluating and Predicting Answer Quality in Community QA. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. ACM, New York, NY, USA, 411–418. https://doi.org/10.1145/1835449.1835518

[54] Milad Shokouhi, Ryen White, and Emine Yilmaz. 2015. Anchoring and adjustment in relevance estimation. In *Proceedings of the 38th International ACM SIGIR Conference on research and development in information retrieval.* 963–966.

[55] Camelia Simoiu, Chiraag Sumanth, Alok Mysore, and Sharad Goel. 2019. Studying the "Wisdom of Crowds" at Scale. In *The Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP-19).* AAAI, 171–179.

[56] G. Stoddard. 2015. Popularity Dynamics and Intrinsic Quality in Reddit and Hacker News. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media.* 416–425.

[57] J. Surowiecki. 2005. *The wisdom of crowds.* Anchor, New York.

[58] Jerry O Talton III, Krishna Dusad, Konstantinos Koiliaris, and Ranjitha S Kumar. 2019. How do People Sort by Ratings?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–10.

[59] Philip E. Tetlock, Barbara A. Mellers, and J. Peter Scoblic. 2017. Bringing probability judgments into policy debates via forecasting tournaments. *Science* 355, 6324 (2017), 481–483. https://doi.org/10.1126/science.aal3147 arXiv:https://science.sciencemag.org/content/355/6324/481.full.pdf

[60] Lyle Ungar, Barbara Mellers, Ville Satopää, Philip Tetlock, and Jon Baron. 2012. *The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions.* Technical Report. 37–42 pages.

[61] Arnout van de Rijt, Soong Moon Kang, Michael Restivo, and Akshay Patil. 2014. Field experiments of success-breeds-success dynamics. *Proceedings of the National Academy of Sciences* 111, 19 (2014), 6934–6939. https://doi.org/10.1073/pnas.1316836111 arXiv:https://www.pnas.org/content/111/19/6934.full.pdf

[62] D. J. Watts. 2012. *Everything Is Obvious: How Common Sense Fails Us.* Random House LLC.

[63] Ryen White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval.* 3–12.

[64] S. S. Wilks. 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Statist.* 9, 1 (1938), 60–62.

[65] Y. Yao, Hanghang Tong, Tao Xie, Leman Akoglu, Feng Xu, and Jian Lu. 2015. Detecting high-quality posts in community question answering sites. *Information Sciences* 302 (2015), 70–82.

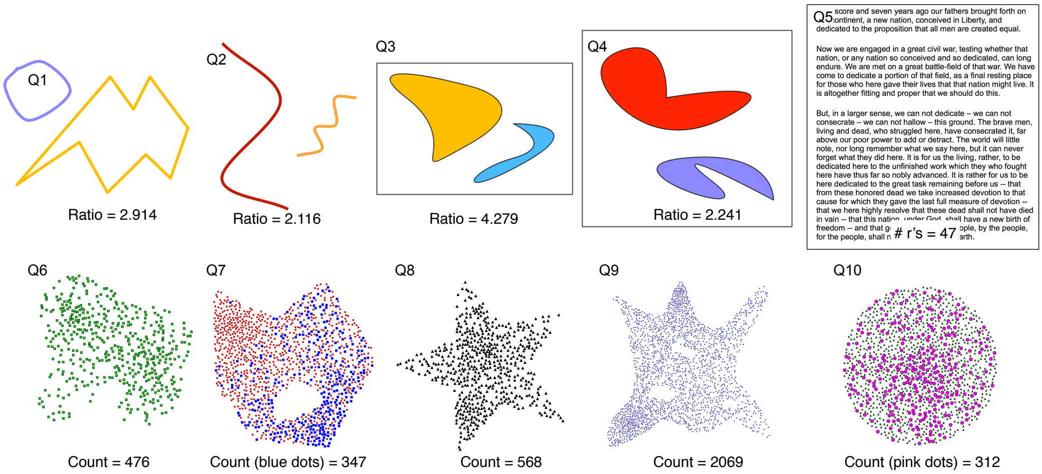
A APPENDIX

In this section, we discuss the experiment questions, the validity of the log-normal guess distribution, alternative answer normalization schemes, and the log-likelihood estimate used to rank answers in simulations. We also discuss the robustness of the simulation results.

A.1 Experiment Details

Experiment questions are shown in Fig. 8. We see that questions cover a variety of visual problems with numerical answers. Questions include finding ratios of lines or areas, counting the number of "r"s in text, or counting dots. We make sure that guesses cannot be easily measured (e.g., lines are not straight and dots are not evenly distributed). The guesses, median guess value, and true values are shown in Fig. 9. We see that the median values are close to the true values, but there is deviation between the two in Questions Q6–10, which happen to be dot-counting questions. In any case, the log-normal fit can only be approximate, since all guesses are required to be at least 1 and in case of Q5-10 are integers. None the less, we find the log-normal approximation useful for later calculations. We also plot how well the data fits a log-normal distribution in Fig 10. We notice that, for most questions, the fit is reasonable or even very good, especially for questions

Q6–10. An exception is Q5, where the data is highly peaked around the correct answer, 47. This is because people have the ability to count the correct answer for this question, while other answers are much more difficult to infer. That being said, Fig. 2 shows us that the normalized answers are very similar to each other, thus despite the disagreement, results are still qualitatively similar to the log-normal distribution.



score and seven years ago our fathers brought forth on :ontinent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate – we can not consecrate – we can not hallow – this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that g ople, by the people, for the people, shall n arth.

Fig. 8. Figures associated with the questions used in our study. Questions were: (Q1) find the perimeter ratio of the larger to smaller shape, (Q2) find the length ratio of the larger to smaller line, (Q3–Q4) find the area ratio of the larger to smaller shape, (Q5) find the number of "r"'s in Lincoln's Ghettysburg Address, (Q6–Q10) find the number of dots. For Q7 and Q10 we specify the color of the dots to count. Correct answers to these questions are listed below the figure.

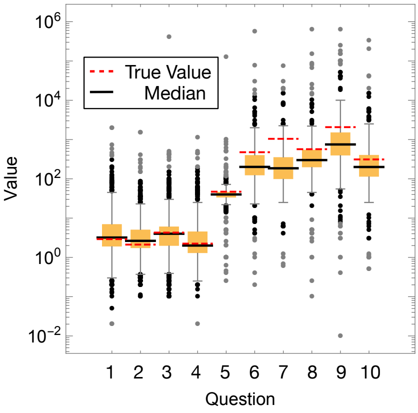


Fig. 9. Guesses for each question. Black line: median guess, red dashed line: true value. We see the median guess are typically close to the true value.

To see if the discrepancy between the guess values and true values affected our model results, we centered answers by $\langle \ln(X) \rangle$ as well as by the true values, as shown in Table 2. We see the main text results (in bold) are very similar regardless of how data is centered. Because we see a

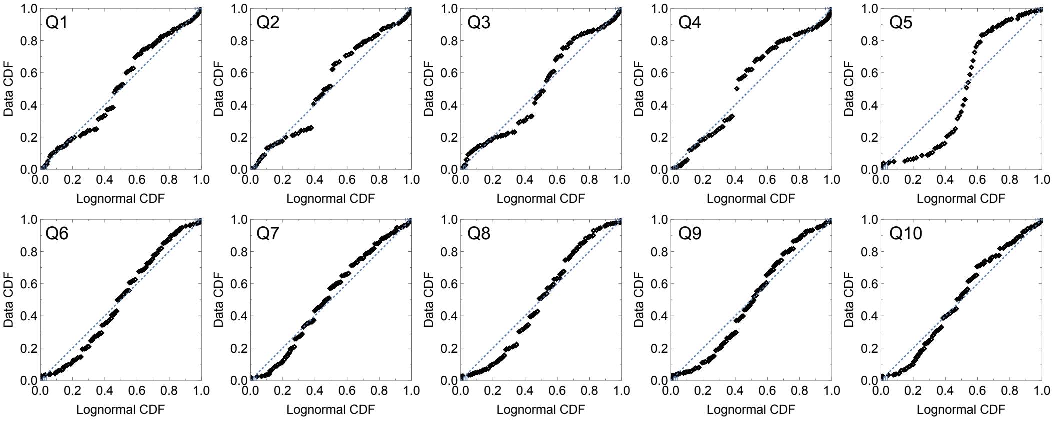


Fig. 10. Quantile-quantile plots comparing empirical data to a log-normal distribution. We see agreement is reasonable except for Q5, and is very good for Q6–10. Q5 corresponds to guessing the number of “r”s in Lincoln’s Ghettysburg Address, which, unlike other questions, can be counted directly.

Table 2. Robustness of Model Fits

Data	Centering	Pr(Matches Data)	Pr(LR_R)	Pr(LR_P)	Pr(LR_0)	$\langle r \rangle$	σ_r	$\langle p \rangle$	σ_p
Control All Q	**Mean**	**0.397**	**0.002**	$< \mathbf{10^{-5}}$	$< \mathbf{10^{-5}}$	**0.28**	**0.02**	**0.05**	**0.02**
Control Q> 5	Mean	0.547	0.023	$< 10^{-5}$	$< 10^{-5}$	0.54	0.05	0.10	0.04
Control Q< 5	Mean	0.61	0.061	$< 10^{-5}$	$< 10^{-5}$	0.19	0.02	0.04	0.02
Control All Q	True	0.546	0.001	$< 10^{-5}$	$< 10^{-5}$	0.33	0.02	0.06	0.02
Control Q> 5	True	0.546	0.020	$< 10^{-5}$	$< 10^{-5}$	0.69	0.04	0.16	0.06
Control Q< 5	True	0.556	0.036	$< 10^{-5}$	$< 10^{-5}$	0.12	0.02	0.03	0.02
Soc. Inf. All Q	**Mean**	**0.472**	$< \mathbf{10^{-5}}$	$< \mathbf{10^{-5}}$	$< \mathbf{10^{-5}}$	**0.08**	**0.02**	**0.21**	**0.01**
Soc. Inf. Q> 5	Mean	0.498	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	0.29	0.06	0.36	0.04
Soc. Inf. Q< 5	Mean	0.52	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	0.06	0.02	0.15	0.02
Soc. Inf. All Q	True	0.504	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	0.11	0.02	0.21	0.01
Soc. Inf. Q > 5	True	0.53	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	0.35	0.05	0.39	0.04
Soc. Inf. Q < 5	True	0.561	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	0.06	0.02	0.14	0.02

Bold results are in the main text. $\langle . \rangle$ represents average and σ represents the standard error.

difference between guesses for dot questions (Q6–10) and ratio question (Q1–4), we separately fit these data subsets to the model. While some of the parameters differed, the qualitative results remain consistent: the model with p and r fit better than simpler models, and p increases in the social influence condition.

A.2 Statistical Methods

A.2.1 Fitting the Decision Model. We fit the decision model defined in the Findings section using maximum likelihood estimation (MLE). This method, however, only provides a point estimate. In order to determine the error of the model parameters, we bootstrap the data (i.e., sample with replacement n times for data of size n) and calculate the MLE values for each parameter. We repeat this step 1000 times to create a parameter distribution, and calculate the standard deviation of this distribution to find parameter error bars.

A.2.2 Comparing Models. The decision model we fit to data has two parameters, however simpler decision models may fit the data equally well. To check whether this is true, we compare the log-likelihood of the two-parameter decision model, $\ell_{\rm DM}$, to a simpler model ℓ_0. Let n be the number of observations, then by Wilks' theorem [64], as $n \longrightarrow \infty$, $\ell_{\rm DM} - \ell_0$ should follow a $\chi^2(\ell_{\rm DM} - \ell_0, k)$ distribution if the data better matches the simpler model, where k is the difference in the degrees of freedom. In our case, the simpler models have one to two fewer degrees of freedom.

A.2.3 Agreement with Data. In order to find out whether our model fits the data well, we compare our model's MLE log-likelihood value to the log-likelihood of data bootstrapped from the model with the same answers (A_1 and A_2) and parameter values (p, and r) as the empirical data. We then say the data agrees with the model if the probability the bootstrapped data fits the model worse than empirical data is greater than 0.1. In practice, we find the probability typically exceeds 0.4, thus the model is consistent with the data. To check the robustness of this agreement method, we also took the MLE of the bootstrapped data by refitting it to the model. Results are virtually identical.

A.3 Calculating Probability Error

To calculate error bars in probabilities, we assume a uniform prior, and use the Beta distribution to create a posterior probability distribution:

$$\Pr(\rho) = \frac{\Gamma(S+F+2)}{\Gamma(S+1)\Gamma(F+1)}\rho^S(1-\rho)^F \tag{11}$$

where S are the number of successes, F are the number of failures, and ρ is the estimated probability of successes. This allows us to calculate error bars for ρ even when $S = 0$ or $F = 0$. In all plots, the point estimation is the MLE: $\hat{\rho} = S/(S+F)$.

A.4 MLE Equation

Let's assume we can independently measure the position bias parameter, p, and random guess parameter, r (e.g. by using results from the present experiment). Let n_t (n_b) and N_t (N_b) be the number of times an answer was chosen when it was ranked first (last) and the total number of votes to all answers when the answer was ranked first (last). Also, recall that if $s(A_{\rm best}, A_{\rm worst}) > 0.5$, $A_{\rm best}$ is the better answer. To estimate $s(A_{\rm best}, A_{\rm worst})$, we first define

$$\Pr(A_{\rm best}\ First|A_{\rm best}, A_{\rm worst}, p, r) = r/2 + (1-r)(p + (1-p)s(A_{\rm best}, A_{\rm worst})), \tag{12}$$

$$\Pr(A_{\rm best}\ Last|A_{\rm best}, A_{\rm worst}, p, r) = r/2 + (1-r)((1-p)(1-s(A_{\rm best}, A_{\rm worst}))), \tag{13}$$

$$\Pr(A_{\rm worst}\ First|A_{\rm best}, A_{\rm worst}, p, r) = r/2 + (1-r)((1-p)s(A_{\rm best}, A_{\rm worst})), \tag{14}$$

and

$$\Pr(A_{\rm worst}\ Last|A_{\rm best}, A_{\rm worst}, p, r) = r/2 + (1-r)(p + (1-p)(1-s(A_{\rm best}, A_{\rm worst}))), \tag{15}$$

where $A_{\rm best}$ and $A_{\rm worst}$ are the respective answers, and "A_x $First(Last)$" means that answer is ordered first (last). The variable $s(A_{\rm best}, A_{\rm worst})$ is the only unknown. Surprisingly, we can infer $s(A_{\rm best}, A_{\rm worst})$ without having to normalize answers, let alone know the answer distribution. The likelihood function of the model is:

$$\begin{aligned} &L(n_t, N_t, n_b, N_b, s, p, r) \\ &= \Pr(A_1\ First|s,p,r)^{n_t}\ \Pr(A_1\ Last|s,p,r)^{N_t - n_t}\Pr(A_2\ First|s,p,r)^{n_b}\Pr(A_2\ Last|s,p,r)^{N_b - n_b} \end{aligned} \tag{16}$$

The log-likelihood function is therefore

$$\begin{aligned} \ell(n_t, N_t, n_b, N_b, s, p, r) = n_t\ln(\Pr(A_1\ First|s,p,r)) + (N_t - n_t)\ln(\Pr(A_1\ Last|s,p,r)) \\ + n_b\ln(\Pr(A_2\ First|s,p,r)) + (N_b - n_b)\ln(\Pr(A_2\ Last|s,p,r)) \end{aligned} \tag{17}$$

To find the MLE of s, we find the solution to

$$\frac{\partial}{\partial s}\ell(n_t, N_t, n_b, N_b, s, p, r) = 0 \tag{18}$$

and then solve for s. The MLE value of $s(A_{\text{best}}, A_{\text{worst}})$ can easily be solved numerically. As long as a researcher records n_t, n_b, N_t, and N_b, we can accurately infer answer quality.

A.5 Simulation Robustness

Simulations in the main text are for the case where the quality ranking algorithm correctly assumes $p = 0.2$ and $r = 0.09$. We explore what happens if one or both assumptions are wrong. For example, we show in Fig. 11 the case when quality ranking assumes $r = 0.0$ when $r = 0.09$. Comparing to Fig. 6, we see results are quantitatively very similar. This is intuitive as randomly choosing the first or last answer with equal probability should not substantially affect their relative ranking.

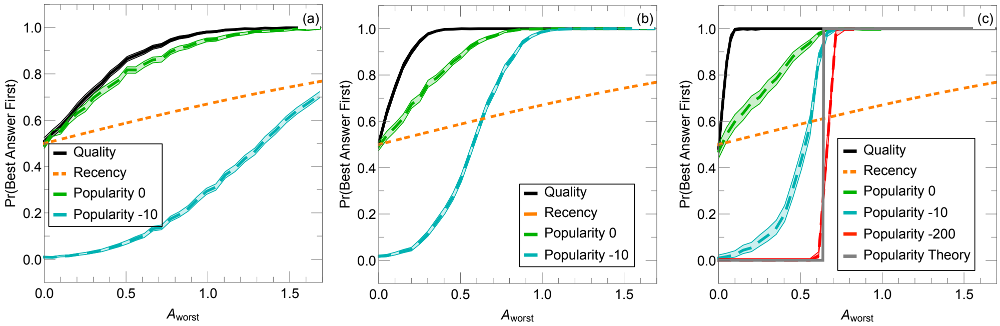


Fig. 11. Comparison of ranking policies when the quality ranking method incorrectly assumes $r = 0.0$ when $r = 0.09$. Plots show the probability the best answer is in first place after (a) 50, (b) 500, and (c) 20,000 votes, when answers are ranked by quality, recency, and popularity. "Popularity 0", "Popularity -10", and "Popularity -200" means that the worst answer starts with a 0, 10, or 200 vote advantage, respectively. Quality ranking (dashed lines) consistently ranks the best answer first, even when the two answers have similar quality (A_{worst}<0.6). Recency tends to perform worst, while popularity is sensitive to the initial number of votes answers start with. Even when both answers start off with equal votes, however, popularity-based ranking does not guarantee that the best answer rises to the top. In comparison, quality-based ranking improves with the number of votes an answer accumulates.

What about if we incorrectly estimate both r and p? For example, we assume $r = 0$ and $p = 0.2$, but in actuality, $r = 0.09$ and $p = 0.1$, 0.2, or 0.3? As shown in Fig. 2, the correct p value could vary between 0.1 and 0.3 for the social influence condition. Results are shown in Fig. 12. Overall, we find that quality ranking still significantly outperforms popularity-based ranking. The only exception is when answers begin with equal votes and $p = 0.1$. In this case, quality ranking is comparable after 20 votes, and slightly worse after 20K votes. A website designer could create small-scale experiments to better infer p, and after applying a corrected p estimate, they should expect quality ranking to again substantially outperform popularity-based ranking. Overall, even when the quality-ranking algorithm is not correctly parameterized, it still performs rather well, and does not not seem to be very sensitive to the estimate of p or r.

Received January 2020; revised June 2020; accepted July 2020

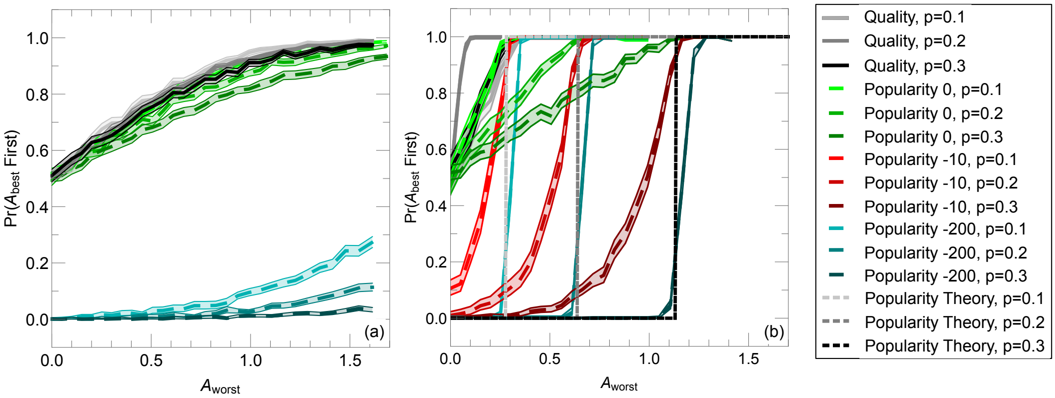


Fig. 12. Ordering answers by quality and popularity after (a) 20 and (b) 20K votes. "Popularity 0", "Popularity -10", and "Popularity -200" means that the worst answer starts with a 0, 10, or 200 vote advantage, respectively. We show different cases when $p = 0.1,\ 0.2$, or 0.3 and $r = 0.09$. In all cases, quality ranking assumes $p = 0.2$ and $r = 0.0$, as a worst-case scenario. Consistent with the main text, quality ranking (dashed lines) consistently ranks the best answer first, and typically outperforms popularity ranking. When $p = 0.1$ and answers begin with equal votes, quality ranking and popularity ranking are comparable.