# How do People Sort by Ratings?

**Jerry O. Talton**
Carta, Inc.
jerry@carta.com

**Krishna Dusad**
University of Illinois at Urbana-Champaign
dusad2@illinois.edu

**Konstantinos Koiliaris**
University of Illinois at Urbana-Champaign
koiliar2@illinois.edu

**Ranjitha S. Kumar**
University of Illinois at Urbana-Champaign
ranjitha@illinois.edu

## ABSTRACT

Sorting items by user rating is a fundamental interaction pattern of the modern Web, used to rank products (Amazon), posts (Reddit), businesses (Yelp), movies (YouTube), and more. To implement this pattern, designers must take in a distribution of ratings for each item and define a sensible total ordering over them. This is a challenging problem, since each distribution is drawn from a distinct sample population, rendering the most straightforward method of sorting — comparing averages — unreliable when the samples are small or of different sizes.

Several statistical orderings for binary ratings have been proposed in the literature (e.g., based on the Wilson score, or Laplace smoothing), each attempting to account for the uncertainty introduced by sampling. In this paper, we study this uncertainty through the lens of human perception, and ask "How do *people* sort by ratings?" In an online study, we collected 48,000 item-ranking pairs from 4,000 crowd workers along with 4,800 rationales, and analyzed the results to understand how users make decisions when comparing rated items. Our results shed light on the cognitive models users employ to choose between rating distributions, which sorts of comparisons are most contentious, and how the presentation of rating information affects users' preferences.

**ACM Reference Format:**
Jerry O. Talton, Krishna Dusad, Konstantinos Koiliaris, and Ranjitha S. Kumar. 2019. How do People Sort by Ratings? In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3290605.3300535


*CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk*

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300535

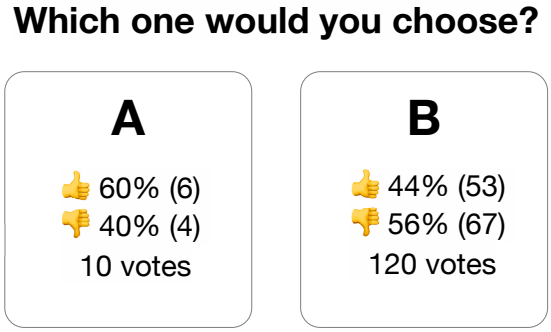


**Figure 1:** This paper studies how people choose between alternatives in a binary thumbs-up/thumbs-down rating system.

## 1 INTRODUCTION

User ratings are ubiquitous across the modern Web. Social and e-commerce platforms crowdsource ratings of content, products, and services at scale, and use these ratings to drive rankings [11], collaborative filtering [19], and recommender systems [17].

The design of user rating systems has been studied in the literature. Sparling et al. [22] compare strategies for presenting rating information, and measure how presentation affects rating distributions, inter-rater reliability, and time-to-rate. Other researchers have studied the effect of prior ratings on users' own decision making [1], modeled the noise that results from erroneous or careless ratings [15], built economic models of user incentives and behavior in rating systems [2], and even used machine learning to predict ratings from content [21].

One of the most common uses of ratings on the Web is to sort a collection of items in user-preference order, for instance to return an ordered list of the highest-rated restaurants or movies in response to a search query. The central challenge of this problem is taking in a distribution of ratings for each item and defining a sensible total ordering over them. This simple problem formulation belies its complexity, since in real-world applications both the distribution of ratings *and* the sample over which that distribution is defined may vary.

The most natural way to compare distributions — by average — is problematic when the sample populations are not both large and of approximately equal size. When the samples are small, the sample average is sensitive to each individual rating and therefore an unreliable predictor of the true population average. When the samples sizes are unequal, quantifying and comparing their relative uncertainties can be problematic.

For the case of binary thumbs-up / thumbs-down ratings, this uncertainty can be estimated mathematically, for instance via Wilson confidence intervals [25], Laplace smoothing [27], or Bayesian formulations [13]. In this paper, we study this uncertainty through the lens of human perception, and ask, "How do *people* sort by ratings?" In particular, we seek to understand the cognitive strategies users employ to choose the more preferable item when presented with two different thumbs-up / thumbs-down distributions [7].

To answer this question, we conducted an online study of 4,000 crowd workers (Figure 1). We partitioned the space of ratings into four representative comparison classes, and collected 48,000 rating-pair preferences over them along with 4,800 rationales justifying users' choices. Patterned off real-world interfaces, we tested three different presentational formats for ratings distributions to better understand how users make decisions when presented with incomplete rating information.

From this data, we determine which kinds of choices are most contentious and time consuming for users, compare users' actual preferences to the predictions made by popular statistical sorting methods, and offer some hypotheses about the cognitive models users employ to make sorting decisions. Our results have implications for the future design and implementation of rating systems.

## 2 RATINGS & RANKINGS

Rankings based on user ratings abound on the Web. Figure 2 shows examples from several highly-trafficked websites employing two of the most popular rating schemes: binary thumbs-up / thumbs-down, and five-star. In this paper we focus on understanding binary ratings, which are easier to reason about since they can be parameterized by only two variables: the number of up-votes $n_u$ and number of down-votes $n_d$.

To rank such distributions, a number of simple formulae have been proposed, such as the *positive difference* $\bar{p} = n_u - n_d$, and the *positive proportion* $\hat{p} = n_u/(n_u + n_d)$ (or *average rating*). The positive difference is not widely used, since it can counterintuitively produce rankings that favor items with a lower percentage of positive votes. The average rating provides a useful summary of the sample distribution, but can be challenging to compare, since proportions elide the size of the population over which they were computed and

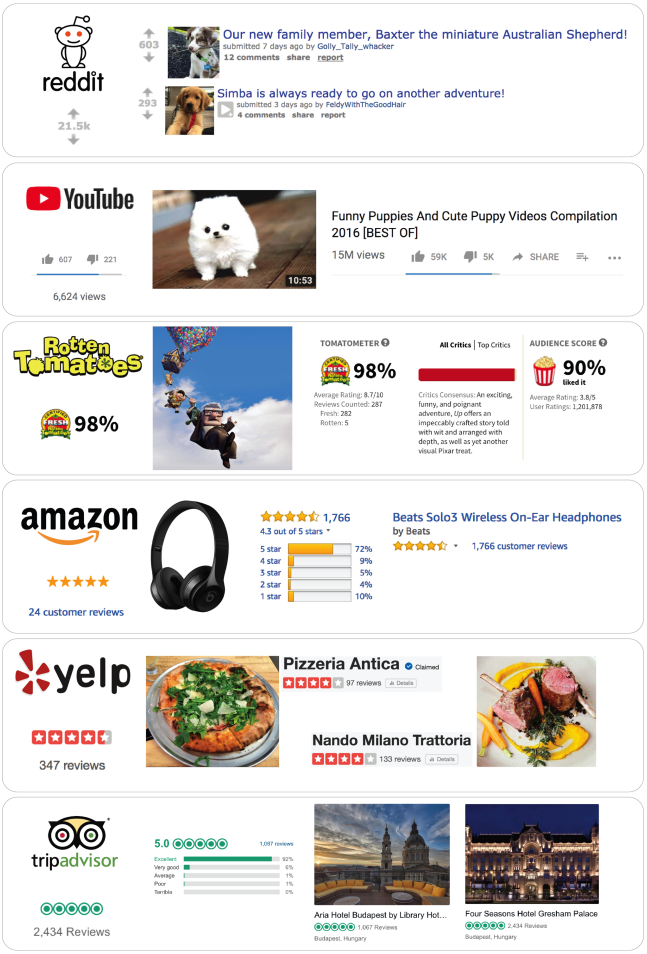


**Figure 2:** Examples of ratings-driven rankings across the Web: Reddit (posts), YouTube (videos), Rotten Tomatoes (movies), Amazon (products), Yelp (businesses), and TripAdvisor (hotels). The top three sites use the binary thumbs-up / thumbs-down ratings we study in this paper; the bottom three employ a five-star system.

there are an infinite number of distributions possessing the same average.

The size of the sample population is material in these comparisons because it is correlated with the uncertainty surrounding the calculated average. When the population is small, the average is sensitive to small changes in the sample. As the population grows large, changing data points fail to move the average much. Difficulty arises when comparing averages between sample populations of differing size: in the example shown in Figure 1, just four additional A downvotes would flip the relative ordering from 60% > 44% to 42.8% < 44%. Four additional B downvotes, in contrast, would change B's positive proportion by little more than one percent. Which, then, is really the "better" choice?

To combat this problem, Miller [12] recommends the *Wilson score* [25] — or more precisely the lower bound of the Wilson confidence interval for a Bernoulli parameter — as the sorting function

$$s_w = \frac{1}{1+\frac{1}{n}z^2}\left[\hat{p} + \frac{1}{2n}z^2 - z\sqrt{\frac{1}{n}\hat{p}\left(1-\hat{p}\right) + \frac{1}{4n^2}z^2}\right],$$

where $n = n_u + n_d$ is the total number of ratings and $z$ is the $(1 - a/2)$-quantile of the standard normal distribution. Like other binomial proportion confidence intervals, the Wilson score provides a lower bound on how far the true population average may lie from the sample average, given some fixed confidence level. As the size of the sample grows, the sampling error shrinks, and the confidence interval converges to the positive proportion from above and below.

Schumacher [20] and Zhang et al. [27] advocate using a different correction to the sample average for ranking via *Laplace smoothing*, which assumes that every item has a single thumbs-up and thumbs-down rating by default $s_\ell = (n_u + 1)/(n_d + n_u + 2)$. In a Bayesian sense, this shrinkage estimator for the positive proportion is equivalent to computing the expected value of the posterior average, given the observed ratings and using a beta distribution with $\alpha = \beta = 1$ as the prior.

## 3 RELATED WORK

The design of user rating systems has been studied in the literature, particularly in the context of online recommendation and reputation systems.

Recommendation engines leverage techniques like collaborative filtering to predict personalized preferences from large collections of user interaction data [19]. *Noise* in this data affects the quality of recommendations, creating an upper bound on prediction accuracy referred to as the "magic barrier" [3]. Researchers characterize this noise as *natural* when it arises from human error or carelessness, and *malicious* when it results from deliberate attacks on the system [15]. Researchers have measured natural noise by examining how consistently users re-rate items [1, 4], demonstrating that removing data contributed by inconsistent users can improve the overall magic barrier of a recommendation system [18]. *Sparse* rating data also poses challenges for recommendation systems: researchers have built economic models of user incentives and behaviors [2], and compared techniques for learning about new users [17].

The choice of rating scale can affect both the quality and quantity of rating data. While finer-grained scales increase time-to-rate, users often prefer them to coarser alternatives, suggesting that increased granularity may reduce noise [1, 22]. Finer-grained scales, however, do not necessarily produce a more discriminative signal. Product ratings collected under a five-star scale often follow a J-shaped distribution, due to purchasing and under-reporting biases: products mostly receive four- and five-stars, some one-star reviews, and almost no scores in the middle [6]. Accordingly, companies like YouTube and Netflix have switched from five-star to binary ratings, and reported significant increases in rating frequency as a result [11]. Kluver et al. [9] developed a information-theoretic framework to model this quality-quantity tradeoff.

With the boom of the sharing economy, platforms like Uber and Airbnb critically rely on users' reviews, building reputation systems to promote trust between strangers. Recent work has studied how information presented on user profiles in such systems can affect perceptions of trust and service quality. Thebault-Spieker et al. found that race- and gender-based profile information did not bias how Mechanical Turk participants rated simulated gig work [23]. Qiu et al. [16] measured how the perception of trust on Airbnb is affected by both the average star rating of a profile and its number of reviews. When presented with a set of profiles that were sufficiently differentiated along these two axes, users placed more trust in an account with roughly ten times more reviews than one with a 1-point higher average rating.

## 4 STUDY OVERVIEW

To understand how people sort by ratings, we conducted two experiments, showing users pairs of ratings distributions and asking them to select the most preferable one. In both experiments, distributions are presented without reference to a particular domain, and denoted by the letters **A** and **B**.

In the first experiment, we aim to understand the cognitive models users employ to make rankings decisions, and identify the most contentious types of comparisons, where people disagree about which distribution is preferable. We partition the comparison space into four categories, sample a set of representative questions from each one, and then measure inter-rater reliability and time-to-rate over a set of crowd workers recruited for task. To understand the cognitive models workers employ to determine their preferences, we collect text rationales for a subset of the comparisons. To reduce the cognitive burden on workers and minimize their mental calculations, we present each distribution fully-specified with the number of up- and down-votes, the positive and negative proportions, and the total number of votes.

In the second experiment, we explore the way users' preferences are influenced by changes in the presentation of rating information. Here, we restrict our investigation to the most contentious category of comparisons from the first experiment, and vary the information architecture of the distribution presentations to mimic popular online sites.

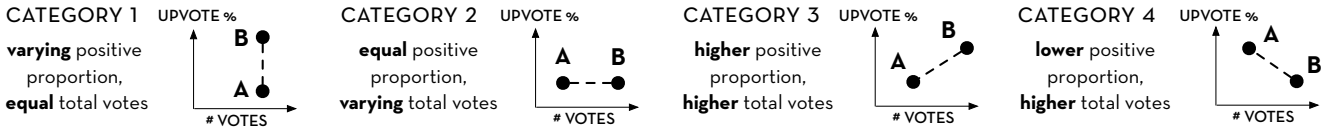


**Figure 3:** The four categories of ratings comparisons we explore in Experiment One.

### Sample Space

To conduct these experiments, we must bound and sample the space of possible comparisons to generate tasks for workers. We can uniquely parameterize each ratings distribution by its positive proportion $\hat{p} \in [0, 1]$ and total number of votes $n \in (1, \infty)$. Generating a comparison between distributions then requires sampling a 4-tuple $(\hat{p}_1, n_1, \hat{p}_2, n_2)$.

To generate a set of informative tasks, we partition the space of comparisons into four semantic categories (Figure 3):

**Category One** comprises comparisons where both distributions have the same number of total votes ($n_1 = n_2$), but the positive proportion varies between them.

**Category Two** comprises comparisons where both distributions have the same positive proportion ($\hat{p}_1 = \hat{p}_2$), but the number of total votes varies between them.

**Category Three** comprises comparisons where one distribution has both a higher positive proportion ($\hat{p}_1 > \hat{p}_2$) and a higher number of total votes ($n_1 > n_2$).

**Category Four** comprises comparisons where one distribution has a higher positive proportion ($\hat{p}_1 > \hat{p}_2$) but a lower number of total votes ($n_1 < n_2$).

To make sampling tractable, we restrict our investigations to distributions with fewer than 10,000 ratings. To ensure adequate coverage over the space of possible comparisons — and generate enough rankings per comparison to observe trends and patterns — we employ a stratified sampling.

First, we linearly partition the range of possible positive proportions into four equal subsets ([0-0.25], [0.25–0.50], [0.50–0.75], and [0.75–1]) and logarithmically partition the range of total ratings ([0–10], [11–100], [101–1,000], and [1,001–10,000]), generating a $4 \times 4$ division of the space of possible ratings distributions. Taking the Cartesian product of this division set with itself yields a $16 \times 16$ partition of the space of possible distribution comparisons.

To generate ranking tasks for a given category, we sample one 4-tuple from each of the 256 partitions of this comparison set, subject to the constraints of the category. Since some partitions do not contain any conforming comparisons for a given category, we omit them from sampling, yielding 40 ranking tasks each for Categories One & Two and 100 for Categories Three & Four. Note that, with the exception of Category One, the logarithmic partitioning of total votes biases the generated comparisons towards order-of-magnitude differences in sample size, which allows us to explore how users navigate large variations in uncertainty.

## 5 EXPERIMENT ONE

To identify contentious comparisons, we recruited a set of online crowd workers through Amazon Mechanical Turk and assigned each one a task containing 14 distribution pairs. Each comparison presented two item cards with labels, a complete description of each ratings distribution, and the query "Which item would you choose?" (Figure 1).

Workers were asked to provide a text rationale for three random comparisons in each task. Once the task was complete, we requested that each participant indicate whether they had any particular domain (e.g., movies, products, restaurants) in mind while making their selections.

### Study Design & Procedure

Our first experiment comprised 2,000 unique tasks, each consisting of one practice comparison, three "sanity-test" comparisons (used for worker validation), two comparisons each from Categories One and Two, and five comparisons each from Categories Three and Four.

The practice comparison — which was fixed across tasks — was intended to introduce the worker to the task and provide examples of acceptable rationales. To prevent malicious or incompetent workers from skewing the collected statistics, we selected three "sanity-test" comparisons from Category One, where both distributions were sampled from large populations and the positive proportion for one clearly dominated the other. Workers who selected the item with lower average rating were asked to redo the task more carefully upon completion. The remaining 14 comparisons were randomly assigned to workers from their respective categories, and presented in random order during the task.

### Measures

To measure contention for each comparison, we compute the inter-rater reliability $I = |n_A - n_B| / n$: the absolute value of the difference between the number of participants who selected item A and the number of participants who selected item B, expressed as a percentage of the total number of respondents. Inter-rater reliability values range from 0% — when participants are equally split between the two choices — to 100% — when the participants are in complete agreement.

We also measure the time-to-rank $T$ each comparison across users, starting from the time the distributions appear and ending when a choice is selected. Since this time data may be noisy (as workers become distracted or leave their

computer mid-task), we track the average time per question on a per-user basis, and exclude times that are more than two standard deviations beyond the worker's mean from our calculations.

We also measure the time-to-rank $T$ each comparison across users, starting from the time the distributions appear and ending when a choice is selected. Since this time data may be noisy (as workers become distracted or leave their computer mid-task), we track the average time per question on a per-user basis, and exclude times that are more than two standard deviations beyond the worker's mean from our calculations.

### Participants

We recruited 2,000 unique Mechanical Turk workers from the US, Canada, the UK, and Australia. Each worker was limited to a single task, and paid $0.17 upon successful completion. After the completion of the study, we paid each worked a retrospective bonus of $0.40 in order to ensure workers earned at least minimum wage for their time.

### Results & Discussion

We collected 28,000 selections for 280 unique comparisons with 2,800 rationales from 2,000 unique workers. The average completion time per task was 4 minutes and 33 seconds. Of the 2,000 participants, 42 (2.1%) failed at least one of the sanity-test comparisons and were asked to repeat the experiment, which all workers did successfully.

We conducted an iterative open coding of the collected rationales to identify some themes behind workers' decisions. Each rationale was independently coded by three members of the research team: rationales without consensus code were discussed until consensus was reached.

In Category One, we observe that people generally prefer the option with higher up-vote percentage when both distributions are sampled from the same size population, selecting it 98.05% of time. This was the least contentious of the four categories, with an average inter-rater reliability of 96.2% across comparisons (Figure 4). Workers spent an average of 5.7 seconds per comparison (min = 4.4, max = 7.7, $s$ = 0.9). Most rationales (59.5%) indicate that workers made their selections by comparing percentages, with a substantial minority (30%) indicating that they compared absolute up-votes. Since the population sizes in Category One are equal within a comparison, these two strategies are functionally equivalent and statistically sound: for a fixed sample size, the distribution with higher positive proportion is more likely to converge to a higher average rating as the sample populations increase.

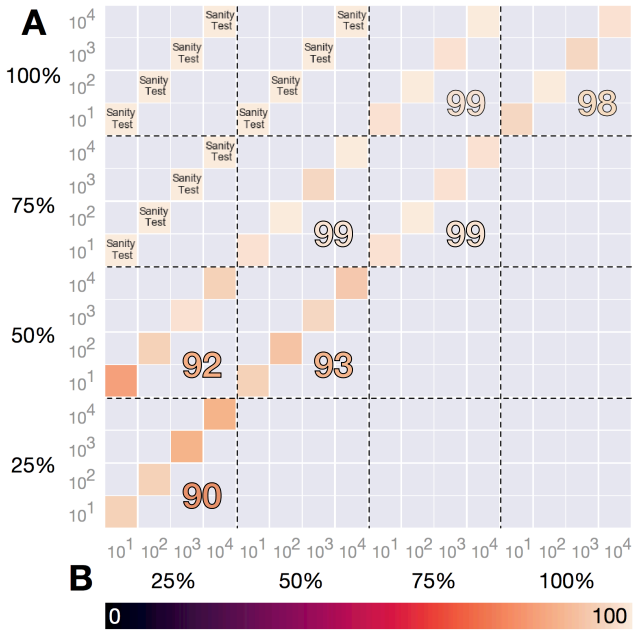


**Figure 4:** The Category One contention heatmap. Each cell is color-coded based on its inter-rater reliability score: the higher the agreement, the lighter the color. The number next to each block represents the average agreement value of each quartile block. Certain comparisons from this category were used as *sanity-tests* in our experimental framework and are marked accordingly.

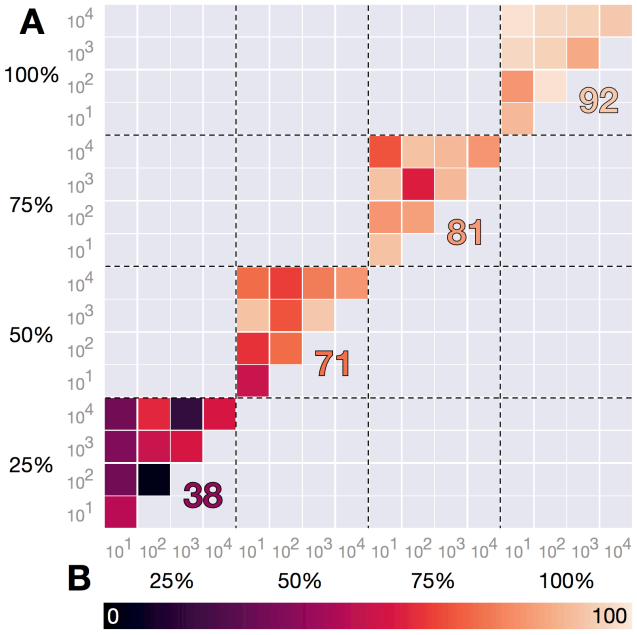


**Figure 5:** The Category Two contention heatmap. Each cell is color-coded based on its inter-rater reliability score: the higher the agreement, the lighter the color. The number next to each block represents the average agreement value of each quartile block.

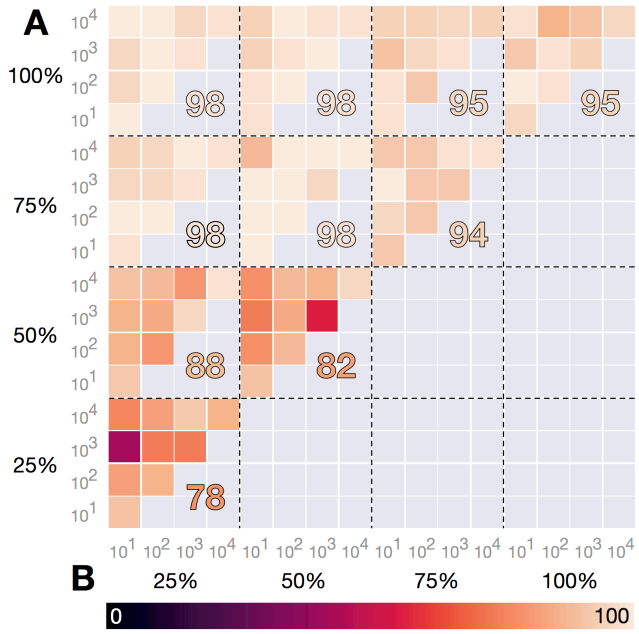


**Figure 6:** The Category Three contention heatmap. Each cell is color-coded based on its inter-rater reliability score: the higher the agreement, the lighter the color. The number next to each block represents the average agreement value of each quartile block.


**Figure 7:** Aggregated time heatmaps for Category Two (left) and Category Three (right), labeled with the average time-to-rate in seconds. For these categories, there is a positive correlation between contention and time-to-rank.

In Category Two, we observe that people prefer the option with higher number of total votes 85.2% of the time when both distributions have the same up-vote percentage, another statistically-sound preference. Contention within this category was also low (Figure 5), with an average $I$ =70.5%. Users report total votes (71.1%) as their primary selection critera, and workers spent an average of 6.8 seconds per comparison (min = 4.9, max = 9.5, $s$ = 1.1)

In Category Three, we see that people prefer the option with higher number of total votes and higher up-vote percentage 96.2% of the time, when one distribution strictly dominates the other on both axes (Figure 6). This, again, is the statistically consistent choice. In this category, contention was low (average $I$ =92.5%) and rationales were split, with 35% reporting up-vote percentage, 14.8% reporting total votes, and 25.1% reporting both factors were responsible for their preferences. This bifurcation of rationales is unsurprising, given that the category involved two varying quantities instead of one. Workers spent an average of 5.8 seconds per comparison (min = 3.4, max = 8.7, $s$ = 1.1).

In all three categories, we observe more contention for comparisons where the positive proportion of both items is less than 50%, with an increase in $I$ of 7.1% in Category One, 32% in Category Two, and 13.9% in Category Three. In this region of the comparison space, users are somewhat *more* likely to choose the item that the evidence suggests is *less* likely to be higher-rated, although the effect is most pronounced in Category Two. This increase in contention is correlated with an increase in time-to-rank (Figure 7), suggesting that users struggle more with these decisions.

There are a number of possible explanations for this behavior. One is through the lens of Tversky and Kahneman's theory of *risk aversion* [8]. When selecting between the better of two promising options, people may implicitly avoid risk to minimize their expected loss: when both options have high ratings, selecting the lower-rated one may feel like "losing the difference" between the averages. On the other hand, when both options seem undesirable, people may actively seek out risk to maximize the potential gain: a choice that seems less *reliably* bad may be preferable to one with a slightly higher average rating.

This explanation is supported by participant rationales. When the positive proportion of both items is less than 50%, we see that 9% of rationales base their decisions on either how *reliable* an item's rating is (e.g., *"It has been used more, so reviews are more reliable", "This is a much larger sample size for A and I feel like it's more reliable than B with only a few votes"*), or on how much *potential* the item has in the future. This latter property is particularly evident when users select items that have very small numbers of votes (e.g., *"Better chance to improve and increase up percentage", "it has 91 votes and it could go up, whereas A is pretty clearly not great."*).

In Category Two, where this preference reversal is most profound, another possible explanation is that the number of total votes may be what Hsee et al. call a "difficult-to-evaluate" attribute in the context of their *evaluability hypothesis* [5]. While people are societally conditioned to compare percentages (e.g., grades in a class) and counts (e.g., home runs batted in a year), understanding the ramifications of population size (the denominator in the positive proportion calculation)

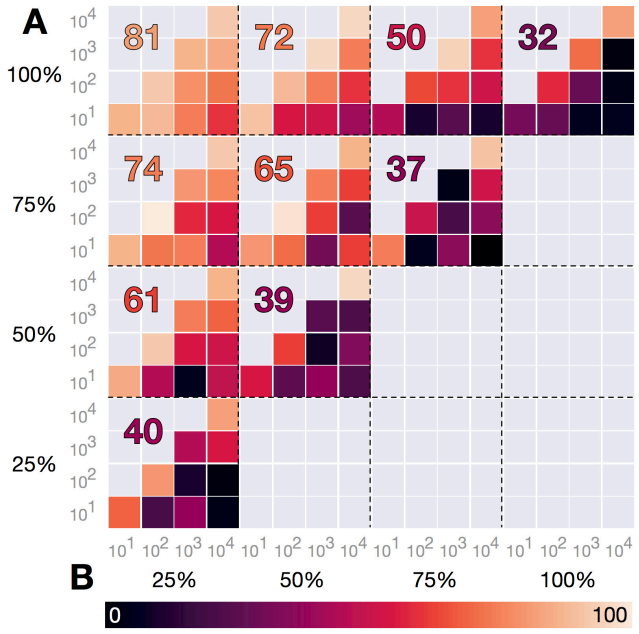


**Figure 8:** The Category Four contention heatmap. Each cell is color-coded based on its inter-rater reliability score: the higher the agreement, the lighter the color. The number next to each block represents the average agreement value of each quartile block.

likely requires more cognitive effort. Under Hsee's hypothesis, such attributes have less influence in decision making.

Category Four — where one distribution has a higher positive proportion but a lower number of total votes — is the most contentious, with average $I$ =55.1% (Figure 8). In these comparisons, users regularly disagree about how to weigh the relative uncertainty between distributions.

While users generally prefer the distribution with higher positive proportion (Figure 9, left), most contention occurs in two regions: when the difference in total votes is high, or when the positive proportions are close. When one distribution has many more total votes but a much lower positive proportion, users are unable to agree whether a "reliable" lower rating is better or worse than an "unreliable" higher one. When the positive proportions are close, users disagree on how to account for differing population sizes in their preferences.

In aggregate, however, users' preferences in this category are still statistically sound, being reliably predicted by Laplace smoothing (75.2%) (Figure 9, right) and the Wilson score (75.0% at a 90% confidence estimate). Workers spent an average of 6.5 seconds per comparison (min = 4.6, max = 9.0, $s$ = 1.0), and the collected rationales for Category Four indicate that up-vote percentage (62.5%), the number of total votes (17.9%), and the absolute number of up-votes (6.2%) all play a role in decision-making.

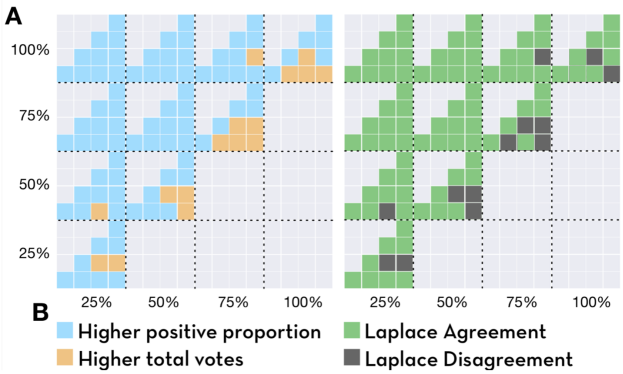


**Figure 9:** In Category Four, a visualization of where users' aggregate preferences agree with positive proportion or higher total votes (left), and the Laplace smoothing ranking (right).

## 6 EXPERIMENT TWO

In order to better understand how the presentation of rating information affects users' preferences, we ran a second experiment on the most contentious comparison category, Category Four. In this experiment, we presented workers with the same distribution comparisons selected in Experiment One, but in two new presentational formats: one showing positive proportion and the number of total votes, and the second showing the absolute number of up- and down-votes (Figure 10).

We chose these particular ratings representations for two reasons. First, they are two of the most widely-used presentational formats for ratings distributions on the Web, mirroring sites like Rotten Tomatoes, YouTube, Amazon, and Yelp. Second, although both formats uniquely specify a binary distribution, they make some potentially useful information that was explicitly presented in Experiment One — the number of up- and down-votes in the first case, and the positive and negative proportions in the second — inaccessible without mental calculation. This allows us to explore whether users invoke Kahneman's System 2 [7] to make ratings comparisons, or merely rely on the information that's readily available.

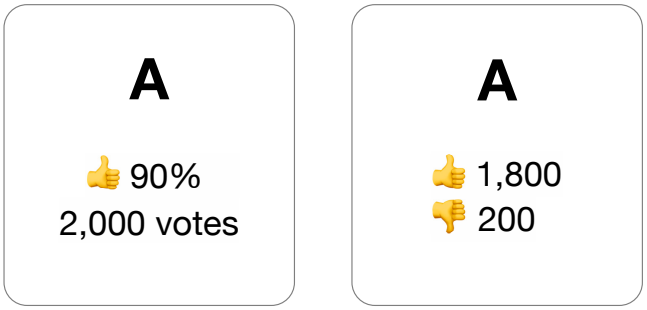


**Figure 10:** The two representations used in Experiment Two: (left) positive proportion and total votes, (right) up-votes and down-votes.

### Study Design & Procedure

We collected 20,000 answers to 200 unique distribution comparisons with a total of 2,000 rationales from 2,000 unique participants, each of whom was paid $0.14 upon successful completion of the task and a $0.40 bonus upon completion of the study. Each task comprised 13 comparisons, and the average completion time per task was 4 minutes and 6 seconds. Out of the 2,000 participants, 54 (2.7%) failed at least one of the sanity-test comparisons and were asked to repeat the experiment, which all workers did successfully. Once again, we conducted an iterative open coding of the collected rationales.

### Results & Discussion

Changing ratings representations changes users' preferences (Figures 11 & 12). We use a two-proportion z-test to determine which rankings are altered between presentation conditions, and observe that 11% of comparisons are significantly different between the complete information and positive proportion conditions, 44% are significantly different between the complete information and up- and down-votes conditions, and 51% are significantly different between the positive proportion and up- and down-votes conditions (at the $p < 0.05$ significance level). Like Kahneman and Tversky's [24]work on the impact of question framing in decision making, here we observe the importance of representation and design in building user-rating ranking systems.

Figure 13 compares inter-rater reliability, average time-to-rate, and agreement with the Laplace smoothing estimate across all three presentational conditions. Observe that, while users take the least time to make decisions in the positive-proportion/total-votes condition ($\bar{x}$=4.8 seconds, min = 3.5, max = 7.0, $s$ = 0.7) they exhibit the highest contention ($I$ =49.2%) and the lowest agreement with Laplace smoothing (72.4%) and the Wilson Score (73.0%), which are robust statistical estimates of the distribution most likely to converge to the highest average rating. Given that the presentation emphasizes positive proportion, it is unsurprising that it is the predominant rationale (57.2%).

In the up- and down-vote condition, users take slightly longer to rank ($\bar{x}$=5.2 seconds, min = 3.4, max = 8.2, $s$ = 1.1)), but with much lower contention ($I$=68.2%) and higher agreement with Laplace smoothing (82.9%) and the Wilson score (76.3%). Although the up-vote percentage is not displayed, it remains the most frequently reported rationale (47%).

One possible explanation for these results is that, when presented with a convenient proxy for decision making — and little additional information (as in the positive-proportion/total-votes condition) — users are content to rely on quicker

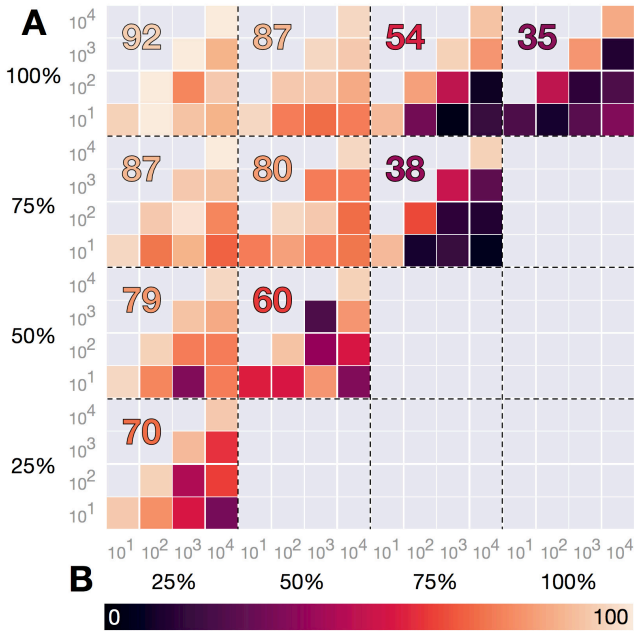


**Figure 11:** The contention heatmap for the up- and down-vote experiment. Each cell is color-coded based on its inter-rater reliability score: the higher the agreement, the lighter the color. The number next to each block represents the average agreement value of each quartile block.

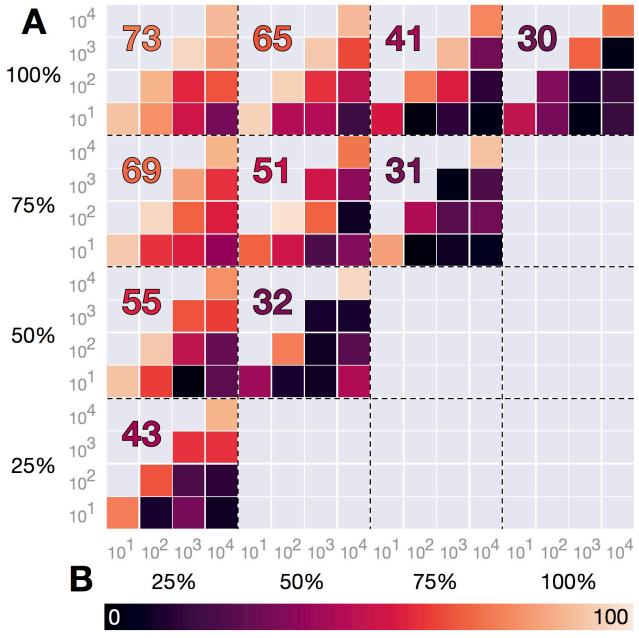


**Figure 12:** The contention heatmap for the positive-proportion/total votes experiment. Each cell is color-coded based on its inter-rater reliability score: the higher the agreement, the lighter the color. The number next to each block represents the average agreement value of each quartile block.

| | Laplace smoothing agreement | Average inter-rater reliability | Average time-to-rate |
|---|---|---|---|
| COMPLETE INFORMATION | 75.2% | 55.1% | 6.5s |
| POSITIVE PROPORTION & TOTAL VOTES | 72.3% | 49.0% | 4.8s |
| UP-VOTES & DOWN-VOTES | 82.9% | 68.2% | 5.2s |

**Figure 13:** Laplace smoothing agreement, average inter-rater reliability, and average time-to-rate for each of the three presentational conditions of Category Four.

System 1 heuristics for making decisions. When complete distribution information is available, users may be confounded by the number of variables and fall prey to *overthinking* [26], resulting in worse decisions in an aggregate statistical sense. When just the right amount of information is given — for instance, enough to suggest that a percentage *should* be computed — System 2 is triggered and users make the soundest judgments.

## 7 DISCUSSION & FUTURE WORK

This paper demonstrates that, when people sort by ratings, in aggregate they make sensible judgments that are statistically sound. When the statistical uncertainty of the distributions being compared is low, users reliably choose the option with the most statistical support. When the positive proportions of both items are less than 50% — or when the difference in total votes is high — the comparisons are more contentious and require more cognitive effort.

These results are consistent with the findings of Qiu et al. [16], who examine Airbnb profiles with four- or five-star average ratings and compare profiles with *low* (1 to 3) and *high* (10 to 50) review counts. If we treat average star ratings as positive proportions, we can map their experiments to one sector of our experimental state space. The majority of participants in our study also prefer the rating with higher total votes, when one rating has a positive proportion of about 0.8 and between 10 and 50 votes, and the other rating has a positive proportion of 1.0 and fewer than 4 votes (Figure 9). This choice, however, is somewhat contentious.

We hypothesize that some of this contention can be ascribed to variations in users' risk models, which may lead to decisions that seem less statistically sound. Users may prefer less "reliable" choices when they feel there is more to gain and less to lose. In the future, incorporating these preferences may improve the performance of Learning to Rank models and personalized recommendation systems [10]: knowing what users most value can help predict how they will choose between products and services in contentious situations. Understanding whether these risk models are more situational or more personal is one interesting avenue for future work. Like prior work on quantifying natural noise in rating systems [1, 4], researchers could measure how consistently users adhere to a risk model by asking them to answer the same set of questions multiple times.

For designers implementing ratings systems online, we make two suggestions. First, in the absence of more sophisticated models, we agree with Schumacher [20] and advocate for the use of Laplace smoothing for ranking. It is a principled, Bayesian estimator; simple to implement; computationally efficient; and, we demonstrate, a reliable predictor of human judgment. Second, since users' preferences are dependent on the presentation of ratings information, we suggest that binary ratings be displayed in up-vote/down-vote format, which minimizes contention and encourages users to make sound judgments.

Future work should also examine whether these findings hold when users express preferences in particular domains. In our post-task surveys, 56.6% of participants reported having a specific domain in mind, such as products (35.9%) or movies (10.8%). It seems likely that users may employ domain-specific risk models, for instance expressing more conservative preferences when buying electronics than when choosing a restaurant for dinner. Similarly, temporal information may play a role in the way people incorporate population size into their judgments: it may seem more acceptable for a newly-opened coffee shop to have a small number of ratings than a book that was released two years ago.

Finally, this paper examines binary ratings distributions. Future work should consider other popular formats, such as the five-star systems employed by Amazon and Yelp. We suspect, however, that this will not be an easy task, since meaningfully sampling these higher-dimensional spaces will require generating orders-of-magnitude more comparisons, and developing statistical uncertainty models to compare them against will necessitate more complex mathematics [14].

## REFERENCES

[1] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. 2003. Is Seeing Believing?: How Recommender System Interfaces Affect Users' Opinions. In *Proc. SIGCHI*. 585–592.

[2] F. Maxwell Harper, Xin Li, Yan Chen, and Joseph A. Konstan. 2005. An Economic Model of User Rating in an Online Recommender System. In *Proc. UM*. 307–316.

[3] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Inf. Syst.* 22 (2004), 5–53.

[4] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. 1995. Recommending and Evaluating Choices in a Virtual Community of Use. In *Proc. CHI*. 194–201.

[5] Christopher K. Hsee, George F. Loewenstein, Sally Blount, and Max H. Bazerman. 1999. Preference reversals between joint and separate evaluation of options: A review and theoretical analysis. *Psychological Bulletin* 125, 5 (1999), 576–590.

[6] Nan Hu, Jie Zhang, and Paul A. Pavlou. 2009. Overcoming the J-shaped Distribution of Product Reviews. *CACM* 52 (2009), 144–147.
[7] Daniel Kahneman. 2011. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York.
[8] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–291.
[9] Daniel Kluver, Tien T. Nguyen, Michael Ekstrand, Shilad Sen, and John Riedl. 2012. How Many Bits Per Rating?. In *Proc. RecSys*. 99–106.
[10] Tie-Yan Liu. 2009. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval* 3, 3 (2009), 225–331.
[11] Nathan McAlone. 2017. The exec who replaced Netflix's 5-star rating system with 'thumbs up, thumbs down' explains why. http://www.businessinsider.com/why-netflix-replaced-its-5-star-rating-system-2017-4
[12] Evan Miller. 2009. How Not To Sort By Average Rating. http://www.evanmiller.org/how-not-to-sort-by-average-rating.html
[13] Evan Miller. 2012. Bayesian Average Ratings. http://www.evanmiller.org/bayesian-average-ratings.html
[14] Evan Miller. 2014. Ranking Items With Star Ratings. http://www.evanmiller.org/how-not-to-sort-by-average-rating.html
[15] Michael P. O'Mahony, Neil J. Hurley, and Guénolé C.M. Silvestre. 2006. Detecting Noise in Recommender System Databases. In *Proc. IUI*. 109–115.
[16] Will Qiu, Palo Parigi, and Bruno Abrahao. 2018. More Stars or More Reviews?. In *Proc. CHI*. 153:1–153:11.
[17] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl. 2002. Getting to Know You: Learning New User Preferences in Recommender Systems. In *Proc. IUI*. 127–134.
[18] Alan Said and Alejandro Bellogín. 2018. Coherence and Inconsistencies in Rating Behavior: Estimating the Magic Barrier of Recommender Systems. *UMUAI* 28 (2018), 97–125.
[19] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based Collaborative Filtering Recommendation Algorithms. In *Proc. WWW*. 285–295.
[20] Aaron Schumacher. 2014. How To Sort By Average Rating. https://planspacedotorg.wordpress.com/2014/08/17/how-to-sort-by-average-rating/
[21] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. How Useful Are Your Comments?: Analyzing and Predicting Youtube Comments and Comment Ratings. In *Proc. WWW*. 891–900.
[22] E. Isaac Sparling and Shilad Sen. 2011. Rating: How Difficult is It?. In *Proc. RecSys*. 149–156.
[23] Jacob Thebault-Spieker, Daniel Kluver, Maximilian A. Klein, Aaron Halfaker, Brent Hecht, Loren Terveen, and Joseph A. Konstan. 2017. Simulation Experiments on (the Absence of) Ratings Bias in Reputation Systems. In *Proc. CSCW*. 101:1–101:25.
[24] Amos Tversky and Daniel Kahneman. 1985. *The Framing of Decisions and the Psychology of Choice.* Springer US, Boston, MA, 25–41.
[25] Edwin B. Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212.
[26] Timothy Wilson and Jonathan Schooler. 1991. Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decisions. *Journal of personality and social psychology* 60 (03 1991), 181–92.
[27] Dell Zhang, Robert Mao, Haitao Li, and Joanne Mao. 2011. How to Count Thumb-Ups and Thumb-Downs: User-Rating Based Ranking of Items from an Axiomatic Perspective. In *Proc ICTIR*. 238–249.