Research Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
2017, Vol. 28(10) 1432–1442
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797617711291
www.psychologicalscience.org/PS
SAGE

The Love of Large Numbers: A Popularity Bias in Consumer Choice

Derek Powell[1], Jingqi Yu[2], Melissa DeWolf[3], and Keith J. Holyoak[3]

[1]Department of Psychology, Stanford University; [2]Department of Psychological and Brain Sciences, Indiana University Bloomington; and [3]Department of Psychology, University of California, Los Angeles

Abstract

Social learning—the ability to learn from observing the decisions of other people and the outcomes of those decisions—is fundamental to human evolutionary and cultural success. The Internet now provides social evidence on an unprecedented scale. However, properly utilizing this evidence requires a capacity for statistical inference. We examined how people's interpretation of online review scores is influenced by the numbers of reviews—a potential indicator both of an item's popularity and of the precision of the average review score. Our task was designed to pit statistical information against social information. We modeled the behavior of an "intuitive statistician" using empirical prior information from millions of reviews posted on Amazon.com and then compared the model's predictions with the behavior of experimental participants. Under certain conditions, people preferred a product with more reviews to one with fewer reviews even though the statistical model indicated that the latter was likely to be of higher quality than the former. Overall, participants' judgments suggested that they failed to make meaningful statistical inferences.

Keywords

popularity, sample size, heuristics, decision making, open data, open materials

Received 10/13/16; Revision accepted 5/2/17

Human adults and children—like chimpanzees, rats, and other mammals—use observations of the behavior of other individuals to help them solve problems and guide their decision making. Social learning, via imitation or emulation, enables the rapid acquisition of knowledge that might be difficult or dangerous to acquire by first-person experience. For example, rats mitigate their individual risk in food selection by making choices based on the selections of their conspecifics and the resulting outcomes (Galef, 2001; Galef & Whiskin, 2000). Similarly, human infants emulate food choices of their caregivers and other trusted partners (e.g., Hamlin & Wynn, 2012; Shutts, Kinzler, McKee, & Spelke, 2009). In this way, social learning enables individuals to "stand on the shoulders of giants"—or at least those of their conspecifics. By aiding problem solving and reducing decision-making risks, social learning has enormous evolutionary and cultural impact (Castro & Toro, 2004; Galef & Laland, 2005; Heyes & Galef, 1996; Tomasello, 2004).

For adults in modern societies, many, if not most, decisions are economic. People must choose which restaurant to frequent, which airline to fly on, and which products and brands to purchase. As is the case for more basic survival decisions, these choices are often guided by social learning: People look to see which goods and services others have chosen and what the results of those decisions have been. For instance, people prefer to buy books that are best sellers (Bikhchandani, Hirshleifer, & Welch, 1998; Chen, 2008) and to download apps with greater download counts (Hanson & Putler, 1996). Marketers are well aware of these facts, and the popularity of a good is often explicitly advertised (Bearden & Etzel, 1982). Indeed, highlighting popular consensus—especially of in-group

Corresponding Author:
Derek Powell, Department of Psychology, Stanford University, Jordan Hall, Building 01-420, 450 Serra Mall, Stanford, CA 94305
E-mail: derekpowell@stanford.edu

members—is a powerful instrument of persuasion across domains (Cialdini, 2009).

Of course, the true power of social learning is not in simply observing how other individuals choose, but also in observing the *outcomes* that result from those choices. The Internet, and especially the rise of consumer-generated content, such as online reviews and testimonials, has yielded an exponential increase in the availability of this sort of highly informative social evidence. One can learn in detail about the outcomes of others' decisions by reading their reviews and can also learn more generally from average scores. However, making use of this information demands additional skills: notably, the ability to make intuitive statistical inferences from summary data, such as average review scores, and to integrate summary data with prior knowledge about the distribution of review scores across products.

In this article, we consider whether people's social learning abilities are sufficiently sophisticated to take advantage of the social evidence in current environments. A large body of research examining heuristics and biases (e.g., Tversky & Kahneman, 1971; see Kahneman, 2011) has shown that people often make decisions using simplified representations or processes, which lead them to exhibit systematic biases. These biases are perhaps especially prevalent in economic contexts involving numerical quantities (for a review, see Griffin, Gonzalez, Koehler, & Gilovich, 2012). Judgment and decision-making researchers have catalogued a variety of errors, including ratio bias (e.g., Kirkpatrick & Epstein, 1992), denominator neglect (e.g., Reyna & Brainerd, 2008), sample-size bias (e.g., Smith & Price, 2010), and numerical anchoring (e.g., Oppenheimer, LeBoeuf, & Brewer, 2008), all of which converge in demonstrating that human judges and decision makers are often poorly calibrated when faced with statistical cues such as means and sample sizes. Still, despite their failures in many statistical reasoning tasks, people are also capable of making statistically optimal judgments across a variety of domains (e.g., Griffiths & Tenenbaum, 2006; Xu & Tenenbaum, 2007).

For online consumers, a typical summary presentation of review information for an item gives cues both to the outcomes of purchases of that item (in the average score) and to the popularity of the item (in the number of reviews it has received). In this context, people might favor more-reviewed items because they view a product's popularity as an important social cue to its quality. However, there might also be statistically driven motivations behind choosing more-reviewed products: In accord with the well-known law of large numbers, a score estimated from a greater number of reviews should be more reliable and give greater certainty about the quality of the product. Consumers may act as intuitive statisticians, preferring more-reviewed products because a larger sample size yields greater certainty regarding product quality. Thus, review counts might provide both social and statistical information.

We investigated the roles of social and statistical inference in product selection in three studies. First, we examined a data set of approximately 15 million Amazon.com reviews to establish the empirical distribution of review scores and the relationship between review count and average review score. We then performed two experiments designed to tease apart how people use the statistical and social information provided by review counts. Participants were presented with pairs of products and asked to select one item from each pair for purchase. Within each trial, one of the items had a relatively large number of reviews, and the other had relatively few reviews. Across trials, we manipulated the difference in review scores between the two products, as well as the overall quality of the pairs.

Average ratings and numbers of reviews might be treated as traditional statistical quantities, supporting statistical inferences. Alternatively, review counts might be treated as explicit social cues about other individuals' choices or behaviors, supporting varieties of social inference. We formalized these different interpretations in two alternative models: a Bayesian model of statistical inference and a heuristic cue-weighting model (Meehl, 1954) of social inference. Our Bayesian model described the choices of an "intuitive statistician," straightforwardly interpreting review scores and counts as statistical quantities and integrating them with prior knowledge to make selections. In contrast, the heuristic model treated number of reviews as a measure of popularity, weighting this cue additively with review score.[1]

For the binary choices presented in our experiments, a social-information account of popularity predicts a bias toward selecting more-reviewed products that is independent of other factors. In contrast, intuitive statisticians are predicted to exhibit a more complex pattern of choice: Popular products should be favored when a high review count supports confidence in high product quality, but should be avoided when a high review count supports confidence in low product quality. That is, preferences for popular choices should be modulated by evidence that outcomes were less than satisfactory.

Analysis of Amazon Review Data

A preliminary question concerned the empirical relationship between product popularity and quality or consumer satisfaction. To examine this relationship, we used data from a total of 15,655,439 reviews of 356,619 products (each with 5 or more reviews) in four product categories: cell phones and accessories, electronics,

Table 1. Summary of the Amazon Review Data

Category	Number of reviews	Number of unique products	Mean score (*SD*)	$\rho_{x,n,\text{ price}}$[a]
Cell phones and accessories	2,989,317	71,746	3.73 (0.712)	.016
Electronics	6,939,859	137,508	3.923 (0.692)	−.006
Health and beauty	2,454,430	65,688	4.09 (0.650)	−.023
Kitchen and dining	3,271,833	81,677	4.10 (0.661)	−.031

[a] This column presents semipartial Spearman rank correlations between average review score (x) and number of reviews (n), controlling for price.

kitchen and dining, and health and beauty products. These data are a subset of Amazon review data collected by McAuley and his colleagues (McAuley, Pandey, & Leskovec, 2015; McAuley, Targett, Shi, & van den Hengel, 2015). Though it seems intuitive that better products should become more popular, research in artificial culture markets has shown that the success of a good is often highly unpredictable and sometimes has only a weak relationship to its quality (Salganik, Dodds, & Watts, 2006).

The relationship between average review score (x) and number of reviews (n), even controlling for price, was negligible within each product category (see Table 1). Figure 1 shows the estimated density of $P(x \mid n)$ across values of n for the four product categories. Conditional probabilities were estimated from the Amazon review data with kernel density estimation, a nonparametric approach to estimating probability distributions from observations. The vertical bands formed by the most probable values within each subplot reveal a generally

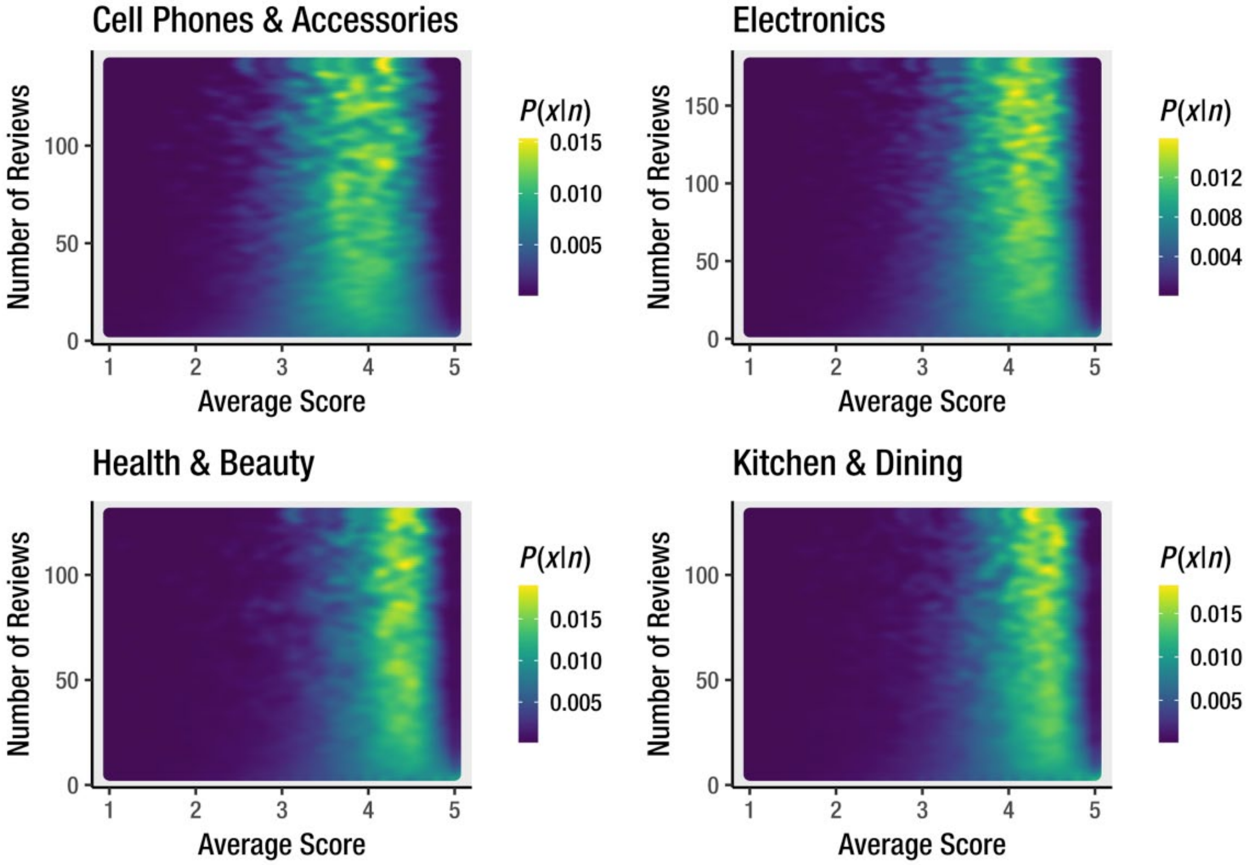


Fig. 1. Results of the analysis of Amazon review data: heat maps showing estimated probabilities of average review score x (on a 5-point scale) across different numbers of reviews (n) for each product category. (For visualization purposes, the plots exclude the top 5% most-reviewed items in each category.)

consistent distribution of review scores across number of reviews (though with somewhat greater variance at low values of n). These empirical results suggest that mere popularity (as indexed by number of reviews) is not a meaningful indicator of product quality. Accordingly, we treated review score as independent of review count in developing our statistical model.

A Bayesian Statistical Model of Product Evaluations

We sought to model behavior of an intuitive statistician. Though there are many ways to infer the quality of a product from a set of online reviews, we attempted to model this task as straightforwardly as possible: as the estimation of a true population mean from a sample mean. We imagined that there is some true value of a given product, θ, and that this true value determines the population of possible reviews for that product. A set of online reviews represents a sample of that population. Thus, we modeled reviews as providing an estimate of the probable true value of a product (θ, taking values ranging from 1 to 5) given the product's mean review score (x) and the number of reviews it received (n), or $P(\theta \mid x, n)$. This problem can be well formulated as one of Bayesian statistical inference. According to Bayes rule, the posterior estimate can be computed by integrating likelihood and prior functions, as follows:

$$P\left(\theta \mid x, n\right) = \frac{P(x, n \mid \theta)P(\theta)}{P(x, n)}.$$

Estimating a population mean using a sample is a straightforward statistical problem. Accordingly, we defined our likelihood function using the hypothetical sampling distributions of the mean for different population values of θ.[2] By the central limit theorem, the sampling distribution of the mean is distributed as $N(\theta, \sigma/\sqrt{n})$. To account for the behavior of means based on small samples, we calculated $P(x,n \mid \theta)$ from the probability density function of a t distribution:

$$P\left(x, n \mid \theta\right) = P\left(x \mid \theta, S\right) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}},$$

where t is equal to $(x - \theta)/(S/\sqrt{n})$, $v = (n - 1)$, and $\Gamma(x)$ is the gamma function.

The review information provided only mean x and sample size n, requiring that S (the standard deviation of the reviews for a given product) be estimated. In accord with the cover story we used in our experiments, we estimated this standard deviation as the average standard deviation across the 71,746 products in the "cell phones and accessories" category (mean S = 1.312).

Finally, we assumed that the distribution of mean reviews provides a reasonable approximation to the distribution of true values θ. Accordingly, we estimated $P(\theta)$ empirically, using the Amazon review data set. As shown in Figure 2, the distribution of mean review scores was heavily skewed in all product categories, with products generally rated to be of good quality. That is, very poor products were rare, good products were common, and excellent products were again rarer. We computed our model estimates using a sampling approach, sampling directly from the empirical prior distribution. In the experiments reported here, we asked participants to make judgments about unknown cell-phone accessories; accordingly, we sampled only from reviews in the "cell phones and accessories" category.

With this statistical model, a selection between a pair of products A and B can be made by comparing the posterior distributions of θ_A and θ_B. Specifically, we calculated $P(\theta_A > \theta_B \mid x_A, n_A, x_B, n_B)$ to determine which product was likely to be superior. One way to interpret this probability is as the predicted probability that participants will choose product A over product B.[3] Figure 3 shows plots of this probability for pairs of products with large and small numbers of reviews.

First, we note the effect of sample size: The rating advantage (or disadvantage) for product A affects the probability that it is superior more strongly when sample sizes are larger. This prediction is quite intuitive, as differences in review scores should matter more when those scores are more precisely estimated by a larger sample.

Second, we note the effect of the absolute quality of the items. As shown in the plots, the model generally favors selection of the more-reviewed product, A (i.e., p(A superior to B) > .50), when its reviews are favorable, but favors selection of the less-reviewed product, B (i.e., p(A superior to B) < .50), when reviews of A are poor. This prediction should be intuitive in light of the prior expectation that a majority of products are of fairly high quality. If mean reviews suggest that one product is of above-average quality and better than another, then as the number of reviews increases, so should one's confidence that the former is indeed the superior product. But if mean reviews are poor (i.e., run counter to prior expectations), then as the number of reviews increases, confidence that the true mean is

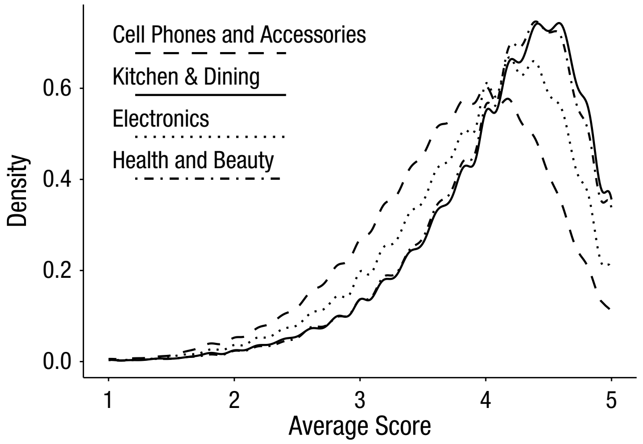


Fig. 2. Kernel density estimation of the distribution of average review scores across the four product categories in the Amazon review data.

indeed low will increase. Consequently, one should prefer the product for which the quantity of poor reviews provides less, rather than more, certainty of its low quality. This prediction is in direct contrast to that of a popularity heuristic, according to which high product popularity is always viewed as favorable.

Experiment 1

Method

Participants. Participants were 138 adults (mean age = 34 years; 60 female, 78 male) recruited from the Amazon Mechanical Turk (MTurk) work-distribution Web site. All participants received $1.00 for participating in the study. We targeted a minimum final sample size of 100 participants to ensure a maximum standard error of .025 when calculating the proportion of participants selecting each product on an individual trial (this maximum standard error resulted from a Bernoulli distribution with a mean of .5).

Materials and design. Experiment 1 consisted of a series of 33 trials, on each of which participants were asked to make a choice between two products. Figure 4 shows an example of a product comparison on a typical trial. Each product was presented with an average star rating (between 1 and 5 stars) and a total number of reviews. In the case of experimental trials, the two products always had different numbers of reviews. The numbers varied slightly from trial to trial, but the difference between paired items was held constant at 125; the more-reviewed product always had approximately 150 reviews, and the less-reviewed product always had approximately 25 reviews. The ratings of the two products were manipulated in a 5 (rating for the more-reviewed product) × 5 (rating advantage for the more-reviewed product) repeated measures design. The ratings for the more-reviewed product ranged across five levels (2.7, 3.1, 3.8, 4.2, and 4.6) that were approximately centered at the average Amazon review score. The rating advantage for that product relative to the less-reviewed product also ranged across five levels: +0.3, +0.1, 0, −0.1, or −0.3. These product ratings and rating advantages were selected as combinations that (according to predictions of the Bayesian statistical model) offered the opportunity to distinguish

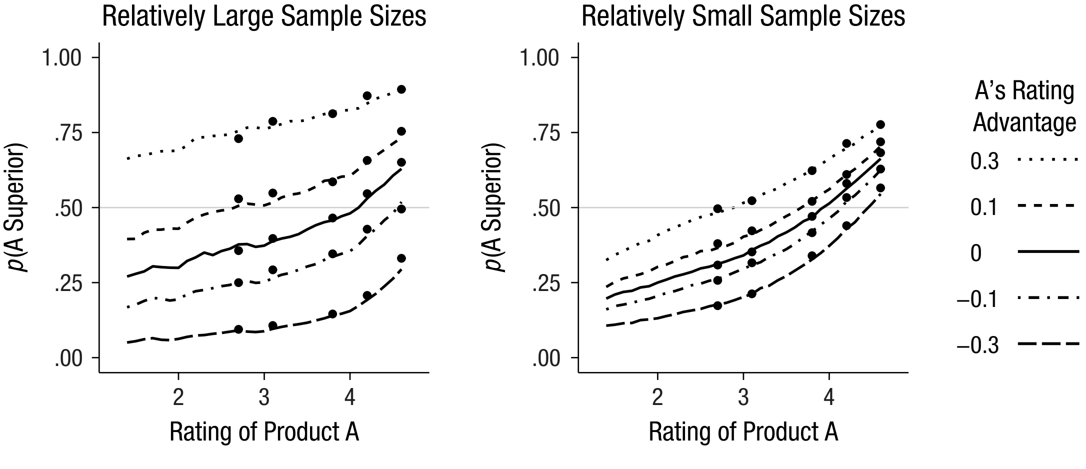
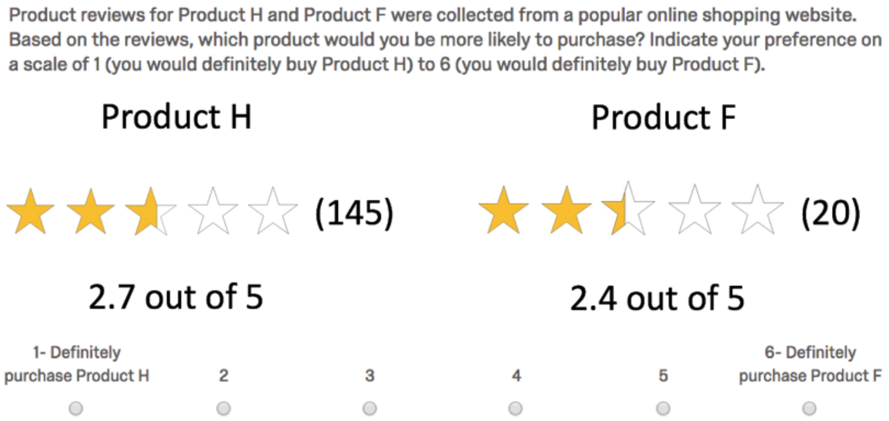


Fig. 3. Plots of the probability that the more-reviewed product, A, is superior to the less-reviewed product, B, as a function of the mean review rating for A. This probability, derived from the Bayesian model, was defined as $P(\theta_A > \theta_B \mid x_A, n_A, x_B, n_B)$. Probability plots are shown for five variations in the difference in ratings between A and B. The graph on the left shows the model's predictions for two relatively large sample sizes (n = 150 for A, 25 for B), used in Experiment 1. The graph on the right shows the predictions for two relatively small sample sizes (n = 26 for A, 6 for B), used in Experiment 2. In both graphs, plotted points indicate the model's predictions for the specific conditions used in Experiments 1 and 2.



Product reviews for Product H and Product F were collected from a popular online shopping website. Based on the reviews, which product would you be more likely to purchase? Indicate your preference on a scale of 1 (you would definitely buy Product H) to 6 (you would definitely buy Product F).

Product H
(145)
2.7 out of 5
Product F
(20)
2.4 out of 5
1- Definitely purchase Product H ○
2 ○
3 ○
4 ○
5 ○
6- Definitely purchase Product F ○

Fig. 4. Display for an experimental trial in Experiment 1.

intuitive statistical inference from a simpler popularity bias. The locations of products on the screen were counterbalanced so that the higher-rated item and the more-reviewed item each appeared equally often on the left and right sides of the screen.

Six trials were filler trials, on which either the two products had equal numbers of reviews (4 trials) or the more-reviewed product had an extreme disadvantage, an average rating 1.9 points below that of the less-reviewed product (2 trials). Finally, we included two check trials, on which the more-reviewed product had an extreme advantage (+1.9). On these trials, it should have been obvious which product was the better choice, so they provided a check on whether participants were paying attention.

Procedure. On each trial, participants' task was to decide which of two different phone cases to purchase. Participants were not given any description of the cases, but were told only that the two products in each pair were similarly priced. The products were arbitrarily labeled with letters from A to Z. Participants were instructed to indicate their preference on a scale from 1 (*would definitely buy the left product*) to 6 (*would definitely buy the right product*). Each participant made judgments for all 33 trials (25 experimental trials + 6 filler trials + 2 check trials).

Results

Check trials. Of the 138 participants, only 6 chose the more poorly reviewed product on one or both of the check trials. These 6 were excluded from further analysis. Thus, data from 132 participants were used in the final analyses.

Experimental trials. First, responses were recoded as binary decisions between the less-reviewed (0) and more-reviewed (1) products. The proportion of participants preferring the more-reviewed item on each trial was calculated (see Fig. 5, top panel).

Participants' decisions differed qualitatively from the predictions of the statistical model (cf. Fig. 5, top panel, with the model predictions in Fig. 3, right panel). Overall, participants showed a far greater preference for the more-reviewed product than the statistical model predicted: In 21 of the 25 conditions, a statistically significant majority of participants chose the more-reviewed product (sign tests, all ps < .01).

This bias was sufficiently strong that participants often favored the more-reviewed product even when the two products had poor quality and the larger number of reviews of the more-reviewed product gave greater certainty of its poor quality. For example, for a pair of products each with an average score of 3.1 stars but one with 29 reviews and the other with 154 reviews, the statistical model yields a .60 probability that the less-reviewed product is superior. Yet for this comparison, more than 90% of human participants chose the more-reviewed product. Excluding cases in which the model was nearly indifferent ($.45 < P < .55$), the model predicted that the less-reviewed product would be the superior choice (i.e., $P(\theta_{\text{more-reviewed}} < \theta_{\text{less-reviewed}}) > .55$) in 11 experimental conditions. Across these 11 conditions, participants performed worse than would be expected by chance guessing (i.e., they preferred the more-reviewed product in 65.5% of trials, $p < .001$ by a sign test), and a majority of participants chose the more-reviewed product despite its being statistically likely to be of lower quality.

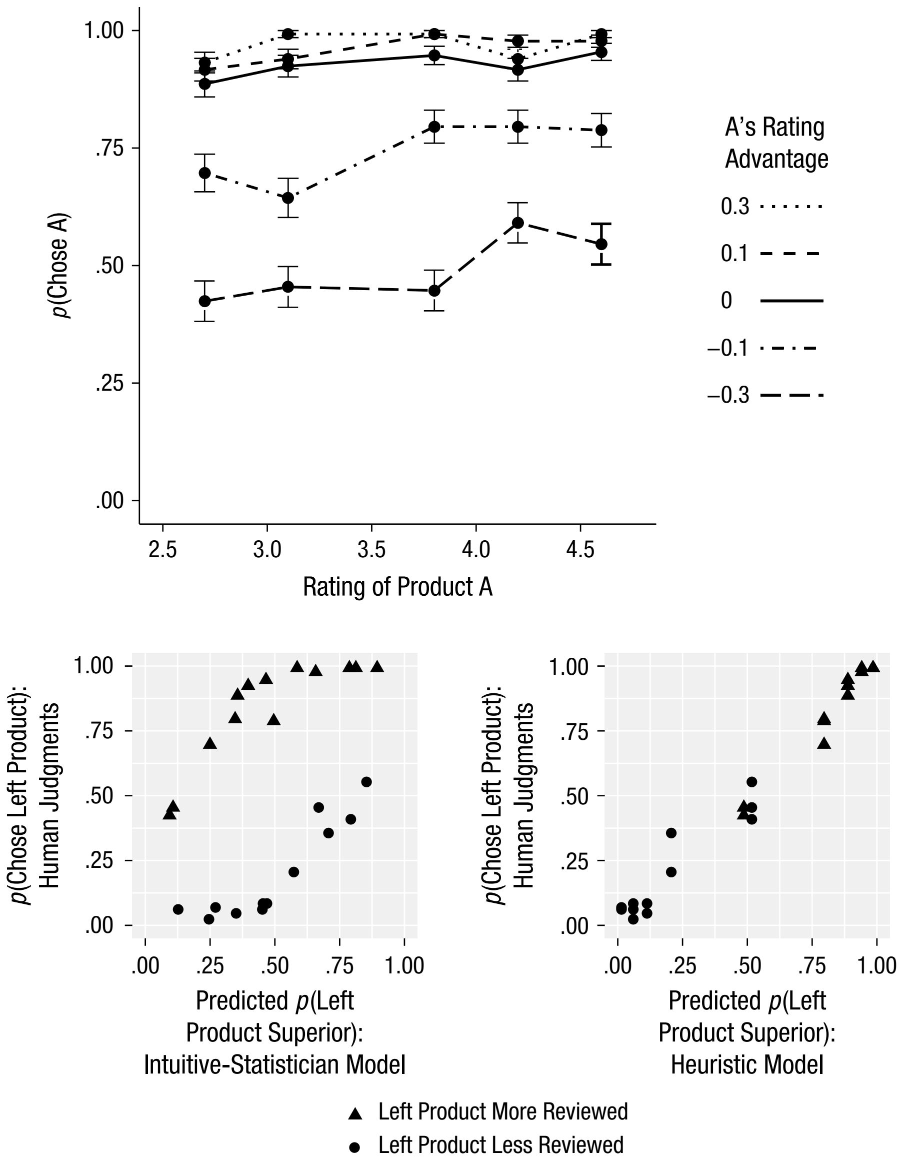


Fig. 5. Results from Experiment 1. The top graph shows the proportion of participants who chose the more-reviewed product as a function of that product's ratings. Results for each level of rating advantage are shown as a separate line. The scatterplots in the bottom row show the observed proportion of participants who chose the product on the left on each trial as a function of the predicted probability of that product being chosen according to the Bayesian statistical model (left) and the heuristic social-inference model (right). The symbols used for the plotted points indicate whether the product on the left was the more-reviewed or less-reviewed item.

Model comparison. For purposes of model comparison, we again treated participants' responses as binary decisions, recoding the data according to whether they chose the product on the left (1) or right (0) side of the display. Figure 5 shows plots of the proportion of participants who chose the product on the left in each experimental trial against the predicted probabilities under the Bayesian statistical model (lower left) and under a multiple logistic regression model (lower right) that combined a predictor for rating advantage and for popularity (based on a binary code indicating which product had more reviews), instantiating a simple cue-weighting heuristic model of the sort long known to be effective in a broad range of prediction tasks (e.g., Meehl, 1954).

As the qualitative analysis presented earlier suggests, the Bayesian statistical model provided a poor fit to the data from Experiment 1 (r^2 = .17, Akaike's information criterion, AIC = 3,352),[4] as did a Bayesian model with uniform priors, *U*(1,5), r^2 = .23, AIC = 3,438. In the bottom left panel of Figure 5, the strong influence of popularity is apparent in the separation of the points in which the left product was the more-reviewed product and the points in which the left product was the less-reviewed product. Participants preferred the left product more strongly than predicted by the statistical model when it was the more-reviewed product, and preferred it less often than predicted when it was the less-reviewed product. In contrast, the social-inference model based on logistic regression provided an almost perfect fit (r^2 = .98, AIC = 203.5). Examining the fitted parameters of the logistic regression model revealed that both rating advantage (*b* = 7.08) and popularity (binary coded, *b* = 4.12) influenced participants' decisions, *p*s < .001.

Experiment 2

In Experiment 1, participants showed a preference for the more-reviewed product even when it was not advantageous to do so. Thus, they apparently failed to use statistical information in estimating the quality of each product. In Experiment 2, we examined how the amount of data available influences participants' decisions for products with fewer reviews than in Experiment 1. If participants are sensitive to the statistical implications of sample size, then they should be less affected by differences in product scores when sample sizes are small, given the greater uncertainty in the mean estimates.

Method

Participants. Participants were 112 adults (mean age = 32 years; 54 female, 58 male) recruited from MTurk. All participants received $1.00 for participating in the study. The target sample size was determined as in Experiment 1.

Materials and design. The design and procedure of Experiment 2 were identical to those of Experiment 1 except that the review sample sizes were reduced to 26 and 6. These values were held constant across all 25 of the experimental trials.

Results

Check trials. Eight of the 113 participants failed one or both check questions. These participants were excluded, and we analyzed the data from the remaining 105 participants.

Experimental trials. As in Experiment 1, participants' responses were recoded to a binary response scale, and the proportion of participants preferring the more-reviewed product was calculated for each trial (see Fig. 6). These responses again departed significantly from the predictions of the Bayesian model (Fig. 3, left panel). As in Experiment 1, participants showed a stronger-than-predicted bias favoring the more-reviewed product; a statistically significant majority preferred the more-reviewed product in 20 out of 25 conditions (sign tests, all *p*s < .01). Also as in Experiment 1, this bias often led participants to make poor choices: Across the 11 experimental conditions in which the more-reviewed product was statistically likely to be of lower quality, participants performed worse than would be expected by chance, preferring the more-reviewed product in 72.3% of trials (*p* < .001 by a sign test).

Model comparison. The observed proportions of choices of the product on the left are plotted against model predictions in the bottom row of Figure 6. As in Experiment 1, a cue-weighting heuristic model based on logistic regression provided a far better fit (r^2 = .97, AIC = 178.6) to the human data than did the predictions of the Bayesian statistical model (empirical priors: r^2 = .05, AIC = 2,105; uniform priors: r^2 = .39, AIC = 2,239). Examining the parameters of the logistic regression model again revealed that judgments were sensitive to both rating advantage (*b* = 6.31) and popularity (binary coded, *b* = 3.95), *p*s < .001.

Meta-analysis

The results of Experiments 1 and 2 were remarkably similar, which is surprising given the large differences in the sample sizes from which review scores were calculated. If participants were at all sensitive to the sample size for reviews, they should have been less sensitive to rating advantage in Experiment 2 than in Experiment 1. We tested this hypothesis by analyzing pooled data from Experiments 1 and 2 in a pair of logistic regression models (employed as data-analysis models). We compared a model predicting choice of the left product from rating advantage and popularity (binary coded) with another model that added a binary experiment variable and an Experiment × Rating Advantage interaction term. These additional variables failed to add to the predictive power of the model (χ^2 = 3.36, *p* = .186). Thus, the meta-analysis indicated that participants were unaffected by the difference in sample sizes between Experiments 1 and 2. These findings suggest that participants failed to engage in statistical inference in any way. Rather, their decisions within each experiment appear to have been based on a heuristic weighting of popularity and average review score.

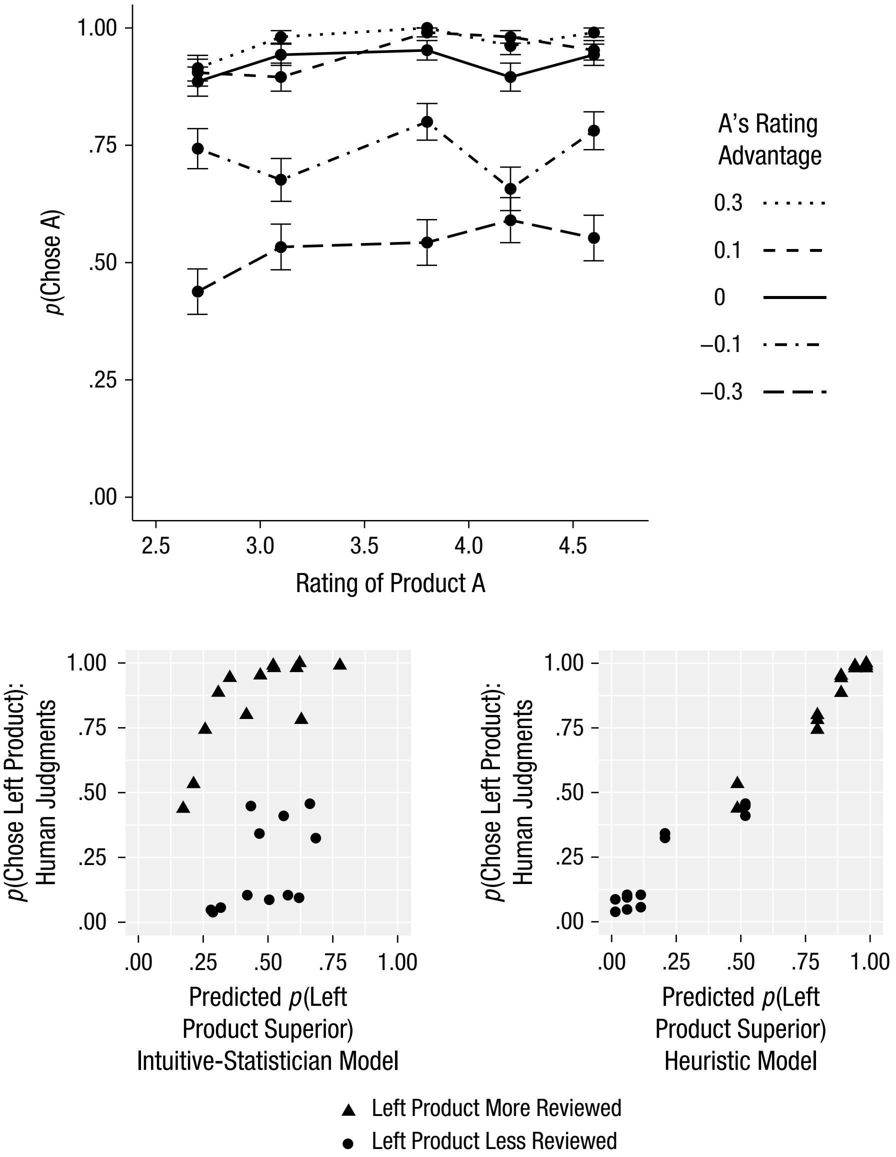


Fig. 6. Results from Experiment 2. The top graph shows the proportion of participants who chose the more-reviewed product on each trial as a function of that product's ratings. Results for each level of rating advantage are shown as a separate line. The scatterplots in the bottom row show the observed proportion of participants who chose the product on the left on each trial as a function of the predicted probability of that product being chosen according to the Bayesian statistical model (left) and the heuristic social-inference model (right). The symbols used for the plotted points indicate whether the product on the left was the more-reviewed or less-reviewed item.

General Discussion

Across two experiments, participants exhibited a strong bias favoring more-reviewed (and thus apparently more-popular) products. In many conditions, participants actually expressed a reliable preference for more-reviewed products even when the larger sample of reviews served to statistically confirm that a poorly rated product was indeed poor. In addition, whereas the Bayesian statistical-inference model predicted that participants should have been considerably less certain in their decisions in Experiment 2 than in Experiment 1, given the smaller sample sizes in Experiment 2, participants were wholly insensitive to the difference in sample sizes across the experiments. Thus, our findings weigh heavily against the Bayesian intuitive statistician as a model of people's product choices. Participants' insensitivity to sample size is particularly troublesome: It is somewhat difficult to imagine a statistical-inference model that would fail to predict differences across such large differences in sample size.

In contrast, participants' responses were well described by a social-inference model weighting cues of popularity (sample size) and differences in average review scores. Our findings suggest that, rather than assessing review scores and sample sizes within a process of statistical inference, participants treated cues about choice outcomes and prevalence as independent and additive factors, without assuming any subtler interaction. That is, they simply weighted these two cues to reach their decision (cf. Meehl, 1954).

These findings suggest that people ascribe outsized importance to the choices of others (as indexed by popularity of items) relative to the outcomes of those choices (as indexed by review score). Participants' bias toward more popular items led them to make decisions that were suboptimal from the perspective of our statistical-inference model. Furthermore, research in artificial culture markets suggests that popularity is a rather weak cue to quality (Salganik et al., 2006), and our own analysis of Amazon review data found no link between product rating and number of reviews. Together, these findings suggest that participants' preference for popular items can be appropriately labeled a bias.

However, it should be noted that the models we have considered are but two among many possible models of this task. Future research might examine alternative models of both statistical and social inference. For example, other statistical-inference models might assume that reviews are drawn from more than one distribution (perhaps distributions of typical and atypical experiences with the product), or that product tokens might vary in their quality (though participants' insensitivity to sample size seems to speak against these possibilities). We can also imagine alternative models of social inference: Perhaps product reviews would be more accurately characterized as arising from an interaction among the features of a product, a user, the user's goals, and so forth. If so, interpreting the desirability of a product could be as much a matter of considering *who* reviewed the product as it is a matter of *what* those reviewers said.

Caveats also apply to the generalizability of our findings. The experimental task we examined represents a rather limited context: a two-alternative forced-choice task with relatively small differences in review scores (−0.3 to +0.3) and relatively large and fixed differences in review counts (an approximately 4:1 ratio or more). Although our findings favor a simple heuristic model of product choice, further research should examine whether other decision contexts might provide evidence that people engage in more sophisticated reasoning when making product choices.

People often fail to make appropriate statistical inferences when presented with raw numbers (e.g., Kahneman & Tversky, 1972, 1973). Future research might investigate whether alternative presentation formats (e.g., sequential presentations or graphical displays) might improve people's use of statistical information (cf. Gigerenzer & Hoffrage, 1995). Such findings might well have practical implications for the format in which information about product reviews is presented online.

One of the many social changes brought about by the advent of the Internet is the proliferation of user-generated content and easy access to social cues from massive groups of people. Greater connectivity has the potential to supercharge one of the most powerful learning mechanisms afforded by evolution and culture. Yet it also places new demands on people's abilities to translate numbers into meaningful social cues. Our findings suggest that these abilities are sometimes limited and unsophisticated, and that it may be beneficial to carefully consider how reviews and other forms of user-generated content are distributed and presented. If people are unable to integrate and apply these cues appropriately in making consequential decisions, this information may do more harm than good. Our findings highlight the power of social cues to guide behavior, but also the relatively simplistic mechanisms by which people sometimes process those cues.

Action Editor

Marc J. Buehner served as action editor for this article.

Author Contributions

D. Powell, J. Yu, and M. DeWolf conceived the project, and J. Yu and M. DeWolf conducted the experiments. D. Powell created the statistical model and analyzed the experimental data. All the authors contributed to early drafts of the manuscript. D. Powell and K. J. Holyoak wrote the final manuscript. K. J. Holyoak supervised the project.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Open Practices

All data, code, and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/7h6cy/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617711291. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at https://www.psychologicalscience.org/publications/badges.

Notes

[1] It should be noted that the distinction between a social-inference model and a statistical model need not be equated with a distinction between heuristic and rational models. Rational social-inference models are also possible, though they are not considered here. Conversely, an inaccurate statistical-inference model might be considered irrational.

[2] Here we made two simplifying assumptions. First, we assumed that number of reviews for a product does not directly provide any information about product quality (an assumption supported by our earlier examination of actual Amazon review data). Second, to apply the central limit theorem, we assumed that individual reviews occur independently of one another.

[3] An alternative method for choosing products in this task is to select the product with the greater expected utility. See our file titled Report.pdf at the Open Science Framework, https://osf.io/7h6cy/, for further discussion of this possibility.

[4] AIC (Akaike, 1974) is an information theoretic index that penalizes model complexity. Lower values indicate better fit.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. doi:10.1109/TAC.1974.110070

Bearden, W. O., & Etzel, M. J. (1982). Reference group influence on product and brand purchase decisions. *Journal of Consumer Research*, *9*, 183–194.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. *Journal of Economic Perspectives*, *12*(3), 151–170. doi:10.1257/jep.12.3.151

Castro, L., & Toro, M. A. (2004). The evolution of culture: From primate social learning to human culture. *Proceedings of the National Academy of Sciences, USA*, *101*, 10235–10240.

Chen, Y. F. (2008). Herd behavior in purchasing books online. *Computers in Human Behavior*, *24*, 1977–1992. doi:10.1016/j.chb.2007.08.004

Cialdini, R. B. (2009). *Influence: Science and practice* (Vol. 4). Boston, MA: Pearson Education.

Galef, B. G., Jr. (2001). Social influences on food choices of Norway rats and mate choices of Japanese quail. *International Journal of Comparative Psychology*, *14*, 1–24. doi:10.1006/appe.2001.0494

Galef, B. G., Jr., & Laland, K. N. (2005). Social learning in animals: Empirical studies and theoretical models. *Bioscience*, *55*, 489–499.

Galef, B. G., Jr., & Whiskin, E. E. (2000). Social influences on the amount of food eaten by Norway rats. *Appetite*, *34*, 327–332.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.

Griffin, D. W., Gonzalez, R., Koehler, D. J., & Gilovich, T. (2012). Judgmental heuristics: A historical overview. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 322–345). New York, NY: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767–773. doi:10.1111/j.1467-9280.2006.01780.x

Hamlin, J. K., & Wynn, K. (2012). Who knows what's good to eat? Infants fail to match the food preferences of antisocial others. *Cognitive Development*, *27*, 227–239. doi:10.1016/j.cogdev.2012.05.005

Hanson, W. A., & Putler, D. S. (1996). Hits and misses: Herd behavior and online product popularity. *Marketing Letters*, *7*, 297–305.

Heyes, C. M. & Galef, B. G., Jr. (1996). *Social learning in animals: The roots of culture*. San Diego, CA: Academic Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Kirkpatrick, L. A., & Epstein, S. (1992). Cognitive-experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology*, *63*, 534–544. doi:10.1037/0022-3514.63.4.534

McAuley, J., Pandey, R., & Leskovec, J. (2015). Inferring networks of substitutable and complementary products. In *Proceedings of the 21th [sic] ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY: ACM.

McAuley, J., Targett, C., Shi, A., & van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *SIGIR 2015: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 43–52). New York, NY: ACM.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Oppenheimer, D. M., LeBoeuf, R. A., & Brewer, N. T. (2008). Anchors aweigh: A demonstration of cross-modality anchoring and magnitude priming. *Cognition*, *106*, 13–26. doi:10.1016/j.cognition.2006.12.008

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*, 89–107. doi:10.1016/j.lindif.2007.03.011

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*, 854–856. doi:10.1126/science.1121066

Shutts, K., Kinzler, K., McKee, K. B., & Spelke, E. S. (2009). Social information guides infants' selection of foods. *Journal of Cognition and Development*, *10*, 1–17.

Smith, A. R., & Price, P. C. (2010). Sample size bias in the estimation of means. *Psychonomic Bulletin & Review*, *17*, 499–503. doi:10.3758/PBR.17.4.499

Tomasello, M. (2004). Learning through others. *Daedalus*, *133*(1), 51–58. doi:10.1162/001152604772746693

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272. doi:10.1037/0033-295X.114.2.245