How endogenous crowd formation undermines the wisdom-of-the-crowd in online ratings

Gaël Le Mens [corresponding author]

Universitat Pompeu Fabra, Department of Economics and Business, Spain

Address: Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain

Email: gael.le-mens@upf.edu

Balázs Kovács

Yale School of Management, Yale University,

New Haven, CT 06511, USA.

Judith Avrahami

The Federmann Center for the Study of Rationality,

The Hebrew University, Jerusalem, Israel

Yaakov Kareev

The Federmann Center for the Study of Rationality,

The Hebrew University, Jerusalem, Israel

Abstract

People frequently consult average ratings on online recommendation platforms before making consumption decisions. Research on the wisdom-of-the-crowd phenomenon suggests that average ratings provide unbiased quality estimates. Yet, we argue that the process by which average ratings are updated creates a systematic bias. In analyses of more than 80 million online ratings, we found that items with high average ratings tend to attract more additional ratings than items with low average ratings. We call this asymmetry in how average ratings are updated 'endogenous crowd formation.' Using computer simulations, we show that it implies the emergence of a negative bias in average ratings. This bias affects particularly strongly items with few ratings, which leads to ranking mistakes. The average-rating ranks of items with few ratings are worse than their quality ranks. We found evidence for the predicted pattern of biases in an experiment and in analyses of large online rating datasets.

[149 words]

Keywords: Sampling bias, Judgment bias, Collective judgment, Wisdom of the crowd

Introduction

People frequently consult online recommendation platforms before making consumption decisions, not only about which item to buy, but also about which book to read, which show to attend, or which hotel to stay at (Nielsen, 2013). Prominently displayed on these websites is the average rating received by the items. Research on the wisdom-of-the-crowd phenomenon suggests that average ratings should be useful to prospective users and consumers: Even if raters have different tastes and information about an item, the average of individual ratings should provide a good estimate of its quality because individual errors cancel each other out (Galton, 1907; Surowiecki, 2005). It has been shown that even small crowds can produce surprisingly accurate aggregate judgments (Budescu & Chen, 2015; Kao & Couzin, 2014; Mannes, Soll, & Larrick, 2014). At the same time, several studies have demonstrated that, on online rating systems, the wisdom-of-the-crowd phenomenon does not always operate as well as one could expect. Ratings are frequently influenced by earlier ratings (Godes & Silva, 2012; Li & Hitt, 2008; Moe & Schweidel, 2012; Muchnik, Aral, & Taylor, 2013; Schlosser, 2005), and sometimes do not faithfully reflect the experience of prior users (Mayzlin, Dover, & Chevalier, 2014). This can lead to a low correlation between collective judgments and quality (Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Salganik, Dodds, & Watts, 2006), which potentially limits the usefulness of average ratings as quality estimates.

While existing research has focused on settings where the crowd is fixed before judgment aggregation, we argue that in many settings the crowd of evaluators for an item is not given exogenously but instead grows *endogenously*. The mechanism we propose is about the extent to which early and non-representative evaluations are corrected by subsequent evaluations. As an illustration, consider a consumer with a highly non-representative taste, who is amongst the first to experience and rate a fairly good restaurant. If this consumer rated the restaurant as ‘extremely good’, others are likely to choose it, experience it, and rate it. The average rating will likely go down and converge to ‘fairly good’. However, if the non-representative consumer rated the restaurant as bad, others are likely to avoid it and the ‘bad’ rating will remain.

In this example, higher average ratings lead more judges to join the set, or 'crowd', of evaluators. Below we demonstrate that such endogenous crowd formation implies the emergence of systematic evaluative biases in average ratings. This pattern of evaluative biases leads, in turn, to systematic ranking mistakes. Suppose that one of two options received more ratings than the other. The average rating of the option with more ratings will likely be higher than the average rating of the option with fewer ratings, even if the latter has higher quality. When there are many options, options with few ratings will suffer a ranking penalty. These evaluative mistakes will emerge even if the social influence mechanisms identified in the prior literature on online rating systems do not operate (Godes & Silva, 2012; Li & Hitt, 2008; Moe & Schweidel, 2012; Muchnik et al., 2013; Schlosser, 2005).

We present our theoretical arguments and empirical tests in three steps. We first build a computer simulation of the dynamics of endogenous crowd formation in the setting of online ratings, through which we delineate the proposed theory and mechanisms. We then report an experiment that tests our theoretical predictions about the emergence of systematic biases in average ratings and the corresponding emergence of ranking mistakes. Finally, we demonstrate the empirical relevance of our theory by reporting analyses of datasets comprising more than 80 million online ratings.

Computational analysis

To demonstrate the effects of the typical mechanism[1] of endogenous crowd formation – in which consumers are likely to choose (and rate) items that already have a high average rating – on the dynamics of average ratings, we contrast it with what happens when the crowd formation is not endogenous or when the opposite dynamics prevails – in which items with *low* average ratings are more likely to be rated.

Model

[1] We test the assumption that this is the case on online rating websites in a later section (Analysis of Field Data 1).

Consider a set of N judges who each evaluate one of two items. Judges arrive sequentially, one in each period. Whether a judge evaluates Item 1 or Item 2 depends on the current average ratings of the two items.

Rating Distribution

Suppose a judge evaluates Item 1 in period t. Let r_{1t} denote this rating. To make it comparable to the typical "star" ratings of online review platforms, we assume it is an integer between 1 and 5. This random variable can be expressed as $r_{1t} = 1 + x_{1t}$, where x_{1t} has a binomial distribution with parameters $(4, p_1)$. The mean of the rating distribution is $q_1 = 1 + 4p_1$. We will refer to q_1 as the (true) quality of Item 1: $E[r_{1t}] = q_1$. Each rating could be higher or lower than the true quality, depending on the specific realization of the random variable (see Fig. S1 for the rating distributions used in the simulations). Ratings for Item 2 follow similar assumptions. We will say that Item 2 has a higher quality than Item 1 when $q_2 > q_1$. Because ratings for an item are independent realizations of a random variable, there is no direct social influence: earlier ratings do not affect the value of subsequent ratings.

Average rating

The average rating for Item i at the beginning of period *t* is denoted by $\hat{S}_{it}$ (this is the unweighted average of the ratings received until the end of period $t - 1$).

Endogenous crowd formation

In each period, a new judge rates an item. The judge decides whether to rate Item 1 or Item 2 as a function of the average ratings of the available items. We assume that the probability that the judge rates Item 1 in period t is given by the following logistic function:

$$\frac{e^{b\,\hat{S}_{1t}}}{e^{b\,\hat{S}_{1t}} + e^{b\,\hat{S}_{2t}}},$$

where b is a parameter that characterizes the sensitivity of the decision to rate Item 1 or 2 to their average ratings. If $b > 0$—the typical case for online ratings—judges are more likely to rate the item with the higher average rating ('positive endogenous crowd formation'). If $b < 0$—the atypical case—judges are more

likely to rate the item with the lower average rating ('negative endogenous crowd formation'). If $b = 0$, the judge is equally likely to rate Item 1 or Item 2, independently of their average ratings; crowd formation is not endogenous.

Model parameters

We report simulations of the model with $N = 10$ judges and with $b = 1$, 0, and -1. The qualities of the two items are $q_1 = 2.8$ and $q_2 = 3.2$. The Supplementary Material (Fig. S1) shows histograms of the corresponding rating distributions.

Results

We denote the values obtained at the end of the simulation period without a 't' index (after the 10 judges have given their ratings to one or the other item). For example, N_1 denotes the number of ratings received by Item 1 by the end of the simulation and $\hat{S}_1$ denotes the final average rating for Item 1. We first consider the cases where average ratings attract additional ratings ($b = 1$).

Overall bias in average ratings

Our main result is that average ratings are negatively biased: $E[\hat{S}_1] = 2.65 < q_1 = 2.8$, $95\%\ CI = [2.64,2.65]$ and $E[\hat{S}_2] = 3.10 < q_2 = 3.2$, $95\%\ CI = [3.10,3.10]$. This pattern is consistent with the prediction of Theorem 1 in Denrell (Denrell, 2005). This is because our model satisfies the premises of this theorem. Just as in Denrell's setting, the negative bias emerges because there is an asymmetry in how the regression to the mean operates. Regression to the mean refers to the statistical phenomenon that in the dynamics of stochastic processes, extreme outcomes—which may reflect error—tend to be followed by less extreme outcomes. Applied to the dynamics of average ratings, it implies that after an additional rating becomes aggregated into the average rating, the average rating will tend to be less extreme than before. High average ratings tend to go down whereas low average ratings tend to go up. If additional ratings tend to arrive when the average rating is high, then downward corrections tend to dominate. Overall, a negative bias in average ratings emerges.

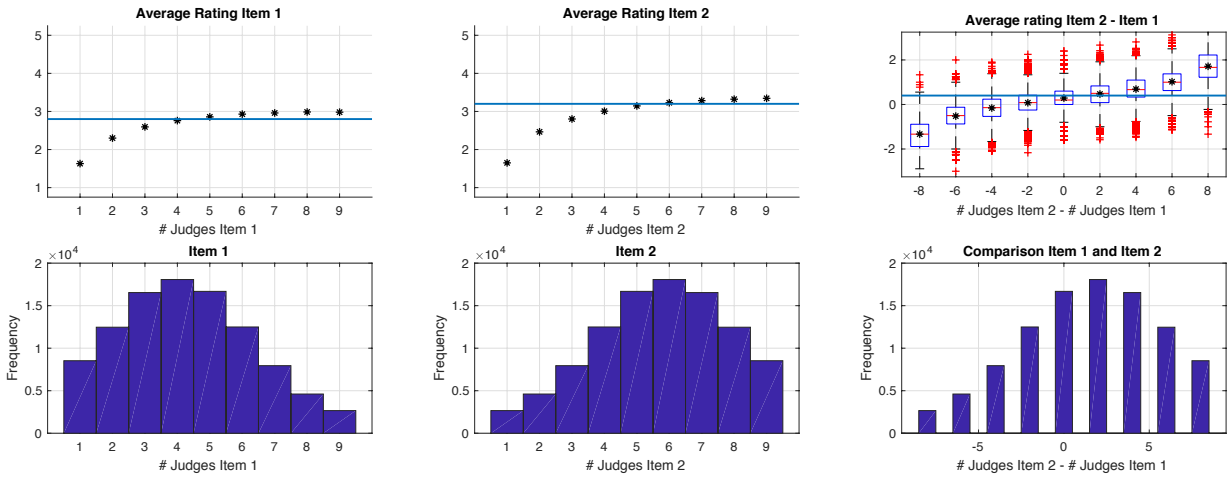

Fig. 1. ***Simulation results with positive endogenous crowd formation (b = 1).*** *On the upper panel plots, the black stars indicate the means, and the horizontal lines indicates the true values (q_1, q_2 and $q_2 - q_1$ from left to right). On the boxplots, the boxes indicate interquartile ranges.*

Systematic pattern of evaluative biases

The negative bias in average ratings is especially strong for average ratings based on few ratings, whereas there is almost no bias for average ratings based on many ratings. This leads to a positive association between number of ratings and average rating. As shown in the top left panel of Fig. 1, the average rating for Item 1, $\hat{S}_1$, increases with the number of ratings obtained by this item, N_1 ($corr(N_1, \hat{S}_1) = .53$). A similar pattern emerges for Item 2 ($corr(N_2, \hat{S}_2) = .52$), as shown on the graph of the top middle panel. The association is strongest for items with few ratings and becomes weaker when the number of ratings increases. This is because as the number of ratings aggregated into the average rating increases, the error decreases, as implied by the law-of-large numbers.

Systematic pattern of ranking mistakes

Remember that Item 2 has a higher quality than Item 1 ($q_1 = 2.8$; $q_2 = 3.2$). The box plot in the top right panel of Fig. 1 shows that systematic *mistakes* emerge: when Item 2 has fewer ratings than Item 1, Item 2 is generally evaluated more negatively than Item 1, even though Item 2 is of higher quality ($q_2 > q_1$). More

precisely, when $N_2 \leq N_1 - 4$, $E[\hat{S}_2] < E[\hat{S}_1]$ and $P[\hat{S}_2 < \hat{S}_1] > P[\hat{S}_2 > \hat{S}_1]$. The histograms of Fig.1 indicate that this case occurs about 15% of the time.

This result about the emergence of collective mistakes seems similar to the collective illusions produced by the model in Denrell & Le Mens (Denrell & Le Mens, 2017). In both models, decision makers are more likely to select an option that has received a higher average evaluation by others. This positive feedback loop leads to the emergence of collective mistakes in both cases. But our model differs from the model by Denrell & Le Mens on two crucial dimensions. First, in our model, the crowd forms endogenously whereas in Denrell & Le Mens, all agents evaluate all the alternatives. Second, whereas repeated evaluations are an essential feature of the model by Denrell & Le Mens, our model works even if agents evaluate the alternatives at most once.

Comparison with other crowd formation mechanisms

As expected, our simulation results show that when crowd formation is not endogenous, average ratings are unbiased estimators of the true quality ($b = 0$, Fig S2A). This result is in contrast with what was obtained with endogenous crowd formation ($b = 1$). This comparison demonstrates that endogenous crowd formation has a causal effect on the emergence of systematic evaluative biases. This happens even when the valence of earlier ratings does not affect the valence of subsequent ratings (i.e., there is no 'social influence bias', as in (Muchnik et al., 2013).

An opposite pattern of results emerges when judges are more likely to evaluate items with low average ratings ('negative endogenous crowd formation', $b = -1$, Fig S2B): there is a positivity bias in average ratings: positive errors are unlikely to be corrected but negative errors will. The positivity bias is stronger for items with few ratings. Finally, when there are fewer ratings for Item 1 than for Item 2, Item 1 is generally evaluated more positively than Item 2 (even though Item 1 has a lower quality than Item 2).

The comparison between the three settings ($b = 1, 0, -1$) demonstrates that the nature of endogenous crowd formation has a causal effect on the emergence of systematic biases in average ratings.

Mean square error

Analyses of the wisdom-of-the crowd phenomenon have frequently focused on the mean square error of the aggregate judgment (MSE) as a measure of accuracy. When there is endogenous crowd formation, the accuracy of the average rating is worse than when there is no endogenous crowd formation. With endogenous crowd formation, the MSE for the average rating of Item 1 is $MSE_{1,b=1} = 0.38$. It is substantially higher than when there is no endogenous crowd formation ($MSE_{1,b=0} = 0.22$). The MSE also increases for Item 2: $MSE_{2,b=1} = 0.28$ versus $MSE_{2,b=0} = 0.22$. The increase is not as large as for Item 1. Since Item 2 has the higher quality, it receives more ratings, and thus judgement aggregation operates better than for Item 1. More crucially for decision making, the mean square error for the difference in average ratings ($\delta\hat{S} = \hat{S}_2 - \hat{S}_1$) is also substantially higher when there is endogenous crowd formation as compared to when there is no endogenous crowd formation ($MSE_{\delta\hat{S},b=1} = 0.69$ versus $MSE_{\delta\hat{S},b=0} = 0.44$).

Robustness checks

In ancillary analyses reported in the Supplementary Material, we analyzed what happens when there is a 'social influence bias' in addition to endogenous crowd formation. We implemented this by assuming that the probability of success in the binomial distribution used to generate the ratings depends not only on the true quality, but also on the current average rating. Simulations show that the social influence bias is not enough to produce on its own a positive association between number of ratings and average ratings (Fig. S3B).

We also analyzed another form of direct social influence in which the number of ratings received by an item affects the valence of subsequent ratings. Existing research on the effect of popularity on evaluations suggests that an association between popularity and average rating could be the result of a positive effect of popularity on rating values (Bikhchandani, Hirshleifer, & Welch, 1992). We call this alternative mechanism the 'direct popularity effect'. We implemented this mechanism by assuming that the probability of success in the binomial distribution used to generate the ratings (see above) depends not only

on the true quality, but also on the number of ratings received by the focal item. Simulations show that the direct popularity effect can produce a positive association between number of ratings and average ratings even if there is no endogenous crowd formation (Fig. S4B). This potentially complicates the interpretation of empirical evidence that indicates a spurious association between number of ratings and average ratings. We return to this issue below.

Finally, additional simulations show that when there are many judges and few items, mistakes almost never occur (see Supplementary Material). This analysis suggests that the evaluative biases we have discussed in this section matter most in settings where many items receive few ratings. We believe it is the case in most online rating environments. For example, in the Amazon.com data we analyze below (78 million ratings of about 8.5 million products), the median number of ratings across products is 2, and 83% of the products have 5 ratings or fewer. In the Yelp.com data (2.2 million ratings of about 77 thousand businesses), the median number of ratings is 8.

Discussion

The results of the simulations show that, when high average ratings attract additional ratings, average ratings become negatively biased. The negative bias is particularly strong for items with few ratings and the asymmetry in the bias leads to ranking errors. This pattern of evaluative biases will emerge even if prior ratings do not influence the valence of subsequent ratings. To test these theoretical findings and to generalize to actual behavior, we conducted an online experiment in which participants rated pictures.

Experiment

As in the simulations reported above, we contrast the effect of the typical endogenous crowd formation, in which high average ratings attract more ratings with the opposite, atypical, crowd formation mechanism in which high average ratings attract fewer ratings. To this end, we compared the dynamics of average rating across 4 independent conditions or 'worlds.' In Worlds 1&2, items with higher average ratings were more likely to be rated and in Worlds 3&4, they were less likely to be rated. We expected that, consistent with

our simulations, a negative evaluative bias would emerge in Worlds 1&2, and that the bias would be stronger for items with fewer ratings than for items with more ratings. Moreover, we expected that items with fewer ratings would suffer a ranking penalty: their average rating ranks would be generally worse than their quality ranks, sometimes severely so. Finally, we expected the opposite patterns to emerge when items with higher average ratings were less likely to be rated (Worlds 3&4).

Design

In addition to the 4 worlds that supported either positive or negative endogenous crowd formation, there were 5 conditions in which the crowd formed non-endogenously (conditions 5_1, 5_2, …, 5_5). These 5 conditions were used to construct unbiased quality estimates that served as a baseline that enabled us to test for the biases that had been predicted to emerge under endogenous crowd formation. In all conditions participants rated 10 pictures.

Manipulating endogenous crowd formation

We asked participants to "have a look at 10 photos and rate them according to your liking." On the experimental screen of worlds 1 to 4, participants did not see the pictures to be rated, but instead saw 50 buttons. On each button was an uninformative label referring to the artist who created that photo ("Artist1," "Artist2," …, "Artist50", see Fig. S5). Participants who clicked on a button were shown the picture on the subsequent screen and had to give it a rating on a scale from 1 to 5 stars (Fig. S6).

Prior research has found that participants in this kind of experiment are more likely to select buttons that are displayed at the top of the screen (Salganik et al., 2006). We used this finding to manipulate the nature of the endogenous crowd formation process. Participants were *randomly assigned* to one of four separate conditions, or 'worlds', as in the MusicLab experiment by Salganik et al. The 4 worlds were initially identical (the 50 pictures were the same and the initial orderings of the buttons were the same), but they were independent from each other in the sense that the order of the items displayed to participants in

World *i* was based on the ratings of the previous participants in that world, hence independent of the ratings of participants in the other worlds.

Worlds 1 and 2 were characterized by an ordering likely to support the emergence of a *positive* endogenous crowd formation: buttons were displayed in *decreasing* order of average ratings by participants who had already rated (the picture with the highest average rating was at the top of the list). The ordering of buttons corresponding to pictures with the same average rating was random. If a picture had not received any rating, the corresponding button was positioned below the buttons of the pictures that had received at least one rating. The two worlds are replications of the same initial setting that might evolve differently. Worlds 3 and 4 had the opposite ordering: buttons were displayed in *increasing* order of average ratings (the button for the picture with the highest average rating was at the bottom of the list).

The initial ordering of the buttons was random and was the same across all 4 worlds. Participants were not told anything about the ordering of the buttons. They were not given any information about the average rating and number of ratings received by a picture. Therefore, the only channel of influence of earlier participants on subsequent participants was the ordering of the buttons on the screen facing the participants.

Estimating 'true' quality

We assume that each picture has a true quality q_i. If each rating of Picture i is a noisy realization of a random variable with mean q_i, the average rating should be close to q_i if it is based on many ratings. To obtain reliable quality estimates, we created 5 conditions in which participants were asked to rate 10 pictures in a random order (without choosing which picture to rate). In condition 5_1, participants rated Pictures 1 to 10. In condition, 5_2, participants rated Pictures 11 to 20, and so on. The number of participants in conditions 5_1 to 5_5 were 51, 52, 50, 50, 49, respectively. We denote by $\hat{q}_i$ the average rating obtained by Picture *i*. We take $\hat{q}_i$ as an independent measure of the quality of a picture (the standard error of the average rating $\hat{q}_i$ as an estimator of the true quality q_i is about 0.2). Because earlier ratings did not affect what happened with later ratings, we refer to these 5 conditions as 'independent' conditions.

We ran the 5 independent conditions at the same time as Worlds 1 to 4. Each participant was randomly assigned to one condition (one of the 4 worlds or one of the 5 independent conditions). Fig. S7 shows the distribution of quality estimates.

Participants

Based on pretesting, we anticipated that with 50 participants per world, the final distribution of number of ratings in the worlds with endogenous crowd formation would be such that the pictures with the fewest ratings would have just 1 or 2 ratings whereas the pictures with the largest number of ratings would have at least 10 to 20 ratings. Therefore, we aimed to recruit 50 participants per world to complete the experiment. The actual numbers of participants were 50, 52, 53 and 47 in Worlds 1 to 4. They were paid \$0.50 for their time. Figure S8 provides histograms of the final number of ratings in each world. Fig. S9 provides histograms of the final average ratings. Fig. S10 provides histograms of the errors in average ratings (the difference between average ratings and the estimated quality) at the end of the experiment. It is important to note that the unit of analysis in this study is not the participant, but the item (the picture). In World 1, one picture (Picture 34) did not receive any rating. This picture is left out of all the analyses of World 1. In the 3 other worlds, all 50 pictures received at least one rating and are thus included in the analyses.

Manipulation check

We first checked that the pictures' ordering indeed influenced participants, that they were more likely to rate pictures displayed closer to the top of the screen (which means higher average rating in Worlds 1&2 and lower average rating in Worlds 3&4).

Consider a participant who was in World j at time t. Let $\hat{S}_{it}$ denote the average rating of Picture i at time t. We estimated logistic regressions that predicted if a participant would rate Picture i as a function of $\hat{S}_{it}$, (with picture fixed effects to account for the initial positions of the pictures). The results are as expected. The effect of the average rating is positive in worlds 1 and 2 and negative in worlds 3 and 4

(World 1: b = 1.50, 95% CI = [1.04, 1.95]; World 2: b = 1.75, 95% CI = [1.30, 2.19]; World 3: b = -0.66, 95% CI = [-0.97, -0.35]; World 4: b = -0.42, 95% CI = [-0.69, -0.1]). In summary, our manipulation of the ordering of the buttons worked as expected: There was positive endogenous crowd formation in World 1&2 and negative endogenous crowd formation in Worlds 3&4.

Predictions

In analyses reported in the SOM (Computational Analyses 2), we simulated the dynamics of average ratings in the experimental setting. Based on these simulations and those of the previous section, we make the following predictions:

- The overall bias in average rating will be *negative* in Worlds 1 and 2 and *positive* in Worlds 3 and 4. Moreover, the average ratings of most pictures will be lower than their qualities in Worlds 1 and 2 and higher than their quality in Worlds 3 and 4.
- There will be systematic associations between number of ratings and average rating. The *negativity* bias will be stronger for pictures with few ratings in Worlds 1 and 2 and the *positivity* bias will be stronger for pictures with few ratings in Worlds 3 and 4.
- There will be systematic associations between number of ratings and ranking mistakes. In Worlds 1 and 2, pictures with few ratings will have average rating ranks worse than their quality ranks whereas pictures with many rating will have average rating ranks better than their quality ranks. The pattern will be opposite in Worlds 3 and 4.

Results

Table 1 summarizes the results. They are consistent with our predictions: The overall bias in average rating was negative in both Worlds 1&2 and positive in Worlds 3&4 (see row 1); this was true for most of the cases in all worlds (see row 2). The association between number of ratings and average rating is positive for Worlds 1&2 and negative for Worlds 3&4 (see row 3) and so is the association between the rank in number of ratings and ranking errors (see row 4).

Table. 1 Experiment Results.

		Endogenous crowd formation process			
		Positive (items with higher average rating receive more additional ratings)		Negative (items with higher average rating receive fewer additional ratings)	
		World 1	World 2	World 3	World 4
Errors in average rating $\varepsilon_i^S = \hat{S}_i - \hat{q}_i$	$E[\varepsilon_i^S]$	-0.51 [-0.69, -0.35]	-0.30 [-0.46, -0.14]	0.26 [0.17, 0.36]	0.36 [0.25, 0.51]
Frequencies: Underestimation / No Error / Overestimation		39/0/10 (N=49)	37/0/13 (N=50)	10/0/40 (N=50)	9/0/41 (N=50)
Association between log10 of the number of ratings and average rating $\varepsilon_i^S = \beta\, log_{10}(1 + N_i) + \gamma + \epsilon_i$,	β	0.82 [0.46, 1.20]	0.83 [0.40, 1.27]).	-0.43 [-0.73, -0.13]	-0.80 [-1.16, -0.45]
Association between ranking error ($\varepsilon_i^{rank} = rank(S_i) - rank(\hat{q}_i)$) and rank in number of ratings $rank(N_i)$: $\varepsilon_i^r = \beta^r\, rank(N_i) + \gamma^r + \epsilon_i$	β^r	0.27 [0.01, 0.53]	0.39 [0.16, 0.63]	-0.30 [-0.49, -0.12]	-0.43 [-0.67, -0.18]

Note: 95% CI in brackets.

For further detail, Figure 2 displays a scatter plot of the average rating as a function of quality; Figure 3 displays a scatter plot of the error in average rating as a function of the number of ratings; Figure 4 displays a scatterplot of the ranking errors in average ratings as a function of rank in number of ratings. On the 3 Figures, each dot corresponds to one picture. The black dots correspond to the pictures for which the final score is outside of the 95% CI for the quality $\hat{q}_i$ (because our measure of quality is based on a finite number of ratings, it approximates the true quality and thus we computed the 95% CI on the true

quality). The grey dots correspond to the pictures for which the final score is within the 95% CI for quality. It is important to note that in all 4 worlds, large errors are very unlikely when the average rating is based on many ratings (Figure 3).

Figure 4 shows that in Worlds 1&2 items with few ratings are likely to have a rank that is worse than their quality ranking. For example, items that were in the lower half of the number of ratings distribution had an average rating rank worse than their quality rank (by 3.8 in World 1 and by 5.1 units in World 2). Conversely, items that were in the upper half of the number of ratings distribution had an average rating rank better than their quality rank (by 3.9 units in World 1 and by 5.5 units in World 2). Note that the increases and decreases in ranks in the two halves are not identical because the distribution of ranking did not allow for an exact 25/25 division of the items; instead the division was made to stay as close as possible to that division. In Worlds 3&4, items that were in the lower half of the number of ratings distribution had an average rating rank better than their quality rank (by 5.1 units in World 3 and by 7.4 units in World 4). Conversely, items that were in higher half of the number of ratings distribution had an average rating rank worse than their quality rank (by 4.3 units in World 3 and by 7.1 units in World 4).

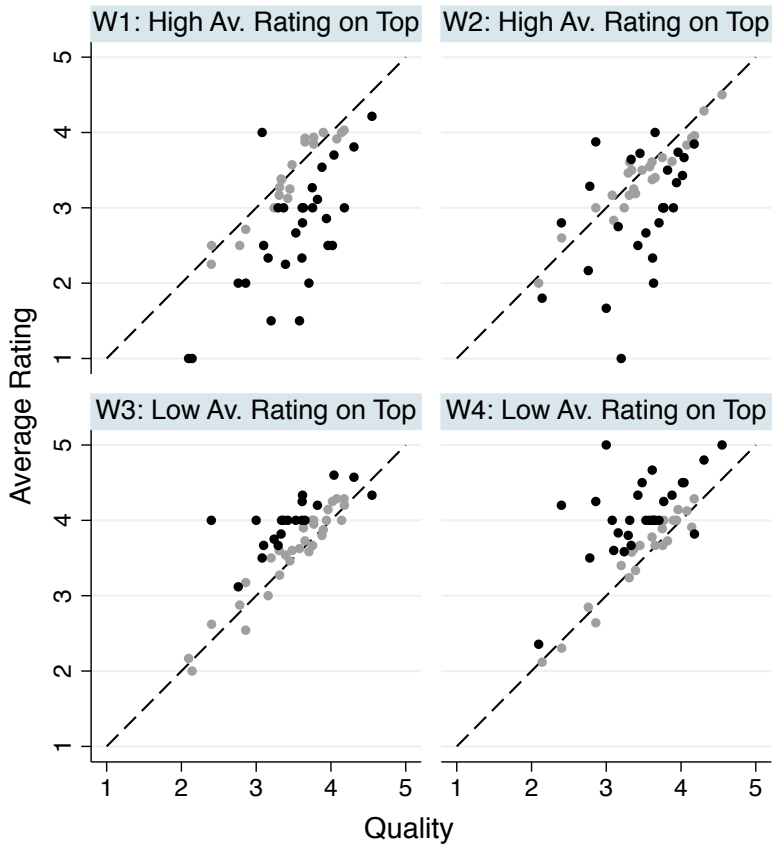

Fig. 2 Experiment. Scatterplot of the average ratings as a function of the quality estimates. Each panel is an experimental World. Each dot represents a picture.

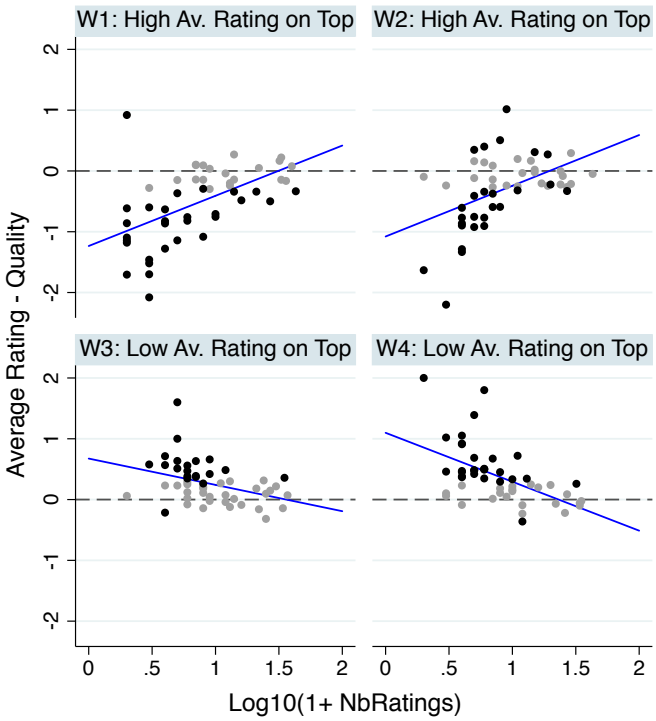

Fig. 3 Experiment. Scatterplot and regression lines of the errors in average ratings as a function of the number of ratings received by a picture.

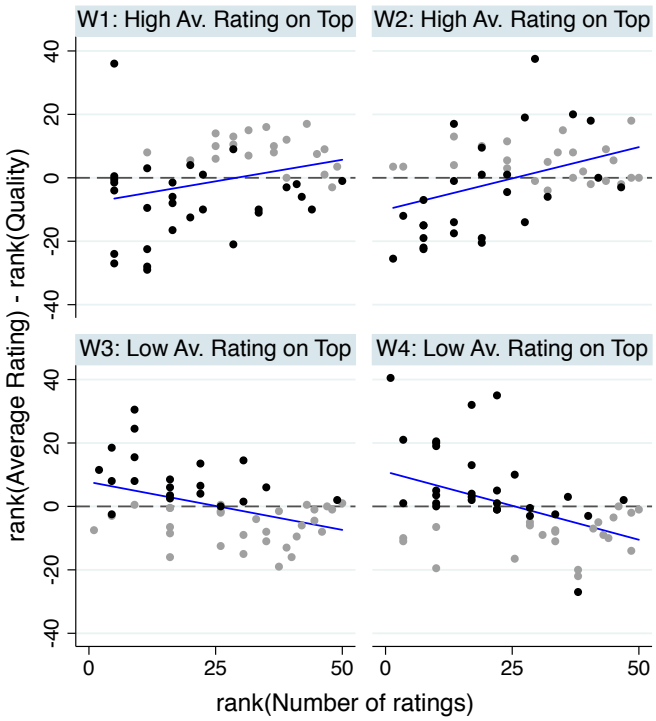

Fig. 4. Experiment. Scatterplot and regression lines of the ranking errors in average ratings as a function of rank in number of ratings.

On the possibility of a 'direct popularity effect'

Note that a 'direct popularity effect' in which the valence of the ratings given by participants is positively affected by the number of ratings, cannot explain our experimental results because the numbers of ratings received by pictures were not available to the participants. Moreover, only the average rating affected the ordering of the buttons, not the number of ratings. Finally, if this mechanism were the dominant explanation for the positive association observed in Worlds 1 and 2, this would imply a similarly positive association in Worlds 3 and 4. But the observed association in Worlds 3 and 4 was negative rather than positive.

Discussion

The results demonstrate that the nature of the endogenous crowd formation process has a causal effect on the emergence of systematic biases in average ratings and of ranking mistakes. Next, we go beyond the controlled, but artificial, setup of an experiment and provide evidence for the assumptions and predictions of our theory from field data.

Analysis of Field Data 1

We demonstrate that endogenous crowd formation, the main assumption of our model, holds in two large scale observational datasets.

Data

We analyzed two datasets: a dataset of 78 million product ratings from Amazon.com (McAuley, Pandey, & Leskovec, 2015a; McAuley, Targett, Shi, & van den Hengel, 2015b) and a dataset of 2.2 million ratings of local businesses provided by Yelp.com. These are the largest publicly available datasets of online ratings we could identify.

Analytical Approach

Consider an item listed on a review website. The dependent variable is the 'arrival rate' of ratings about the item. Formally, it is defined as $\rho_i(t) = \lim_{\Delta t \to 0} \frac{P[t < \tau_t \leq t + \Delta t]}{\Delta t}$, where τ_t is the earliest time after *t* at which a new rating is posted for item *i*. When $\rho_i(t)$ is high, there are frequent new ratings. When $\rho_i(t)$ is low, new ratings are infrequent. Let $\hat{S}_{i,t}$ denote average rating of item *i* at time *t*. It is defined as the average of the ratings received until *t*.

Next, we specify the arrival rate as a 'duration' model of how $\rho_i(t)$ depends on the current collective evaluation, $\hat{S}_{i,t}$, and other covariates (Blossfeld & Rohwer, 2001). Following the standard approach in the analysis of duration models, we assume that $\rho_i(t)$ is an exponential function of $\hat{S}_{i,t}$ and other covariates. We use the following model specification:

$$\rho_i(t) = exp\left(a + b\hat{S}_{i,t} + cX_{i,t}\right), \quad (1)$$

where $X_{i,t}$ is a vector of covariates and a, b, and c are the coefficients to be estimated by maximum likelihood estimation. Positive endogenous crowd formation corresponds to a positive value for b: in this case, higher average ratings ($\hat{S}_{i,t}$) attract additional ratings. Model specifications include the $\log_{10}$ of the number of ratings received by the focal item and fixed effects for the number of years since the item was first listed on the review website, for the calendar year in which the rating was posted, and for item category (all included in the generic term $X_{i,t}$). Standard errors are clustered by item.

Results

We found that in the Amazon.com data, a 1-point increase in average rating was associated with a 16% increase in the *arrival rate* of new rating instances ($exp(b) = 0.16$ with $b = 0.15$, 95%CI = [0.15, 0.15], Model 1, table S2). In the Yelp.com data, the corresponding effect was 14% ($exp(b) = 0.14$ with $b = 0.13$, 95%CI = [0.12,0.13], Model 2, table S2). In summary, there is a positive effect of average rating on the arrival of rating instances in these two large datasets.

We also estimated equation 1 separately for each product category with at least 100,000 ratings in the Amazon.com data (table S3) and for each business category with at least 10,000 ratings in the Yelp.com data (table S4). The coefficients for all the product categories in the Amazon data are positive and significant. In the Yelp data, this is the case for 38 out of 44 business categories. For one category, the coefficient is positive but not significant ($p > .05$). For 5 categories, the coefficient is negative, (not significant for 3 categories and significant for 2 categories, with $p > .05$). Overall, in almost all categories, there is a positive effect of average rating on the arrival of rating instances.

These results are consistent with the hypothesis that positive endogenous crowd formation occurs on online recommendation systems. They do not provide definitive evidence for a causal effect of average ratings on the flow of additional ratings, however. Our results could possibly be explained by unobserved heterogeneity in item quality. Although quality is generally not observable to users of rating websites, a quality proxy (e.g., item picture or description) could affect both the rating valence and the number of ratings. In this case, the positive estimated beta coefficients would not reflect endogenous crowd formation, but instead an exogenous effect of quality. Our results about the endogeneity of crowd formation should thus be interpreted with caution because it is not possible to completely rule out this alternative mechanism.

Discussion

These results suggest that endogenous crowd formation is a widespread phenomenon on review websites. We address the issue of unobserved heterogeneity in quality and provide stronger evidence for the effect of average ratings in the next study.

Analysis of Field Data 2

We collected an online rating dataset with a replicated structure like the worlds in our experiment: we observed the dynamics of average ratings about the same items in two separate environments. This allowed us to perform a stronger test of the hypothesis that average ratings affect the flow of ratings. It also allowed us to demonstrate the emergence of the predicted systematic evaluative biases in field data.

Data

We obtained product ratings from Amazon.de (Germany) and Amazon.fr (France) of all the iPad cases with at least one rating on both Amazon.fr and Amazon.de (76,195 ratings of 788 products). The crucial feature of our data is that the average ratings and popularity information (number of ratings) displayed on each country website are based on ratings posted *only* on that country's website.

We collected data on iPad cases because these products are commodity items and substitute to each other. We conjectured that there would be little cultural difference in appreciation for these products across countries. This intuition was confirmed by the high correlation between the average ratings on the two countries' websites when average ratings are based on many individual judgments and are thus relatively reliable quality estimates (e.g., $\rho = 0.75$, $N = 73$, for items that received at least 30 ratings on both websites) (Fig. S12).

Analytical Approach – Endogenous Crowd Formation

Although the data have a structure like our experimental data, there is an important technical difference: evaluators did not come one after each other to provide a fixed number of ratings. It is thus not possible to analyze the data using a discrete time framework where each evaluation is a 'period'. Instead, ratings could occur at any moment. Therefore, we focus our attention not on the judge but on the item as the unit of analysis. We use a continuous time estimation framework. The dependent variable is the arrival rate of judgments about an item *i* on a country website by unit of time. It is the number of judgments posted about Item *i* on that website per unit of time.

We aim to demonstrate a positive dependence of the arrival rate on the collective evaluation in the focal country, conditional on quality. Quality is unobserved, which makes this demonstration generally difficult. To address this difficulty, we estimated conditional logit regressions that predicted whether a judgment occurred on the French or German website. This approach matches the data about item *i* from the two country websites (Chamberlain, 1979; Hosmer, Lemeshow, & Sturdivant, 2013). Under the assumption

that the effect of (unobserved) quality on the flow of ratings is the same in the two countries, we can specify a model in which quality disappears from the estimation equation.

Suppose that the arrival rate of ratings on each country website is expressed as the exponential of a linear function, as in the previous section (eq. 1). Let q_i denote product quality, $\hat{S}_{i,t}^{FR}$ and $\hat{S}_{i,t}^{DE}$ denote the average ratings of product *i* on the French and German websites at time *t*, $X_{i,t}^{FR}$ and $X_{i,t}^{DE}$ denote a set of covariates pertaining to the French and German websites, and a^{FR} and a^{DE} denote country-specific terms (that capture baseline differences in propensities to post ratings between France and Germany). We have:

$$\rho_i^{FR}(t) = exp\left(a^{FR} + aq_i + b\,\hat{S}_{i,t}^{FR} + cX_{i,t}^{FR}\right), \quad (2)$$

$$\rho_i^{DE}(t) = exp\left(a^{DE} + aq_i + b\,\hat{S}_{i,t}^{DE} + cX_{i,t}^{DE}\right), \quad (3)$$

where a^{FR}, a^{DE}, a, b, and c are real-valued coefficients. If $b > 0$, there is positive endogenous crowd formation: higher average ratings attract additional ratings.

Suppose that product *i* receives a rating at time *t*. The probability that this occurs on the German website is given by

$$\frac{\rho_i^{DE}(t)}{\rho_i^{DE}(t) + \rho_i^{FR}(t)} = \frac{1}{1 + e^{-(a^{DE} - a^{FR} + b(\hat{S}_{i,t}^{DE} - \hat{S}_{i,t}^{FR}) + c(X_{i,t}^{DE} - X_{i,t}^{FR}))}}. \quad (4)$$

This equation corresponds to the specification of a logistic regression in which the dependent variable is equal to 1 if the judgment occurs on the German website and 0 if the judgment occurs on the French website. The crucial feature of this equation is that product quality does not appear on the right-hand side. Thus, estimated coefficients will not be subject to biases potentially induced by the fact that quality is unobserved.

Results – Endogenous Crowd Formation

The estimated b coefficient is positive ($b = 0.64$, 95% $CI = [0.06, 1.23]$, Model 1 in table S5). This indicates positive endogenous crowd formation. The effect persists when controlling for the number of

ratings (Model 2 in table S5). These results are strong evidence for positive endogenous crowd formation, because these analyses control for heterogeneity in product quality.

Prediction– Relation between number of ratings and the size of the evaluative biases

It is not possible to obtain an estimate of true quality independent of the ratings that form the average ratings we analyze. Therefore, we cannot directly test if average ratings are biased. We can, however, measure the association between number of ratings and error in average ratings by relying on the replicated structure of our dataset. Our theory predicts that in this case a spurious *positive* association between number of ratings and average ratings will emerge (this corresponds to the case $b > 0$ in the simulations reported above).

Analytical Approach– Relation between number of ratings and the size of the evaluative biases

The data compare the same items in two rating environments. This replicated structure controls for the main effect of unobserved heterogeneity in item quality. We estimate linear regressions in which we regress the difference in average ratings between the two countries on the difference in (the $\log_{10}$ of) the number of ratings. Specifically, assuming that the effect of the number of ratings is the same in the two countries, we can write:

$$\varepsilon_i^{FR} = \widehat{S}_i^{FR} - \hat{q}_i = \beta \, log_{10}(1 + N_i^{FR}) + \gamma^{FR} + \epsilon_i^{FR}, \quad (5)$$

and

$$\varepsilon_i^{DE} = \widehat{S}_i^{DE} - \hat{q}_i = \beta \, log_{10}\left(1 + N_i^{DE}\right) + \gamma^{DE} + \epsilon_i^{DE}. \quad (6)$$

In these equations, β characterizes the strength of the association between number of ratings and the bias in average ratings.

Let $\Delta\hat{S}_i = \hat{S}_i^{DE} - \hat{S}_i^{FR}$, $\Delta P_i = log_{10}(1 + N_i^{DE}) - log_{10}(1 + N_i^{FR})$, $\Delta\gamma = \gamma^{DE} - \gamma^{FR}$, $\Delta\epsilon_i = \epsilon_i^{DE} - \epsilon_i^{FR}$. Using equations 5 and 6, we get:

$$\Delta\hat{S}_i = \beta\Delta P_i + \Delta\gamma + \Delta\epsilon_i. \quad (7)$$

By estimating eq. 7, we will obtain an estimate of β even though quality is not observable (see the SM for validation of the method on simulation and experimental data).

Results – Relation between number of ratings and the size of the evaluative biases

Estimations of eq. 7 indicate that a 10-fold increase in the number of ratings implied a 0.35-point increase in average rating ($\beta = 0.35$, 95% CI = [0.23, 0.47], N = 788, see Fig. 5). Additional analyses consisting of linear regressions with product fixed effects confirmed this finding (table S6). Consistent with the pattern in Fig. 1, the association between average rating and number of ratings was strongest at low levels of popularity ($\beta = 0.81$, 95% CI = [0.34, 1.28]) and almost nonexistent when the average rating was based on more than 10 ratings ($\beta = 0.051$, 95% CI = [-0.07, 0.18]) (table S6, Models 3 and 4). Further analyses showed that this *positive* association is unlikely to be explained by a positive effect of popularity on rating values because lagged popularity has a *negative* effect on rating values (table S7), consistent with existing findings of negative trends in ratings (Godes & Silva, 2012; Li & Hitt, 2008; Moe & Schweidel, 2012). In other words, the 'direct popularity effect' is unlikely to have operated in this setting.

A possible limitation of this study is that our analyses can only control for the main effect of unobserved heterogeneity. These analyses cannot control of the effect of potential interactions in which an unobserved cause leads some products to receive more favorable and more numerous ratings in one of the two countries. This could be the case if some products were the object of an effective advertising campaign in one country but not in the other country. Similarly, it could be that some products were liked more in one of the two countries because they fitted better with the local taste (reviewers on the French and German websites are not randomly allocated to country websites). As discussed in the data description, we believe this to be unlikely, because iPad cases (the items in our data) are highly substitutable commodity items and we cannot think of any good reason why users of the German and French websites might have systematically different tastes for some types of iPad cases.

Discussion

We found further evidence that crowd formation on rating websites is endogenous. We also found evidence that the negative bias is stronger for items with few ratings (as in Worlds 1&2 in the Experiment). These empirical results should, however, be interpreted with caution because of the absence of random assignment of potential evaluators into the two rating environments. Nevertheless, they parallel and complement the findings of our experiment.

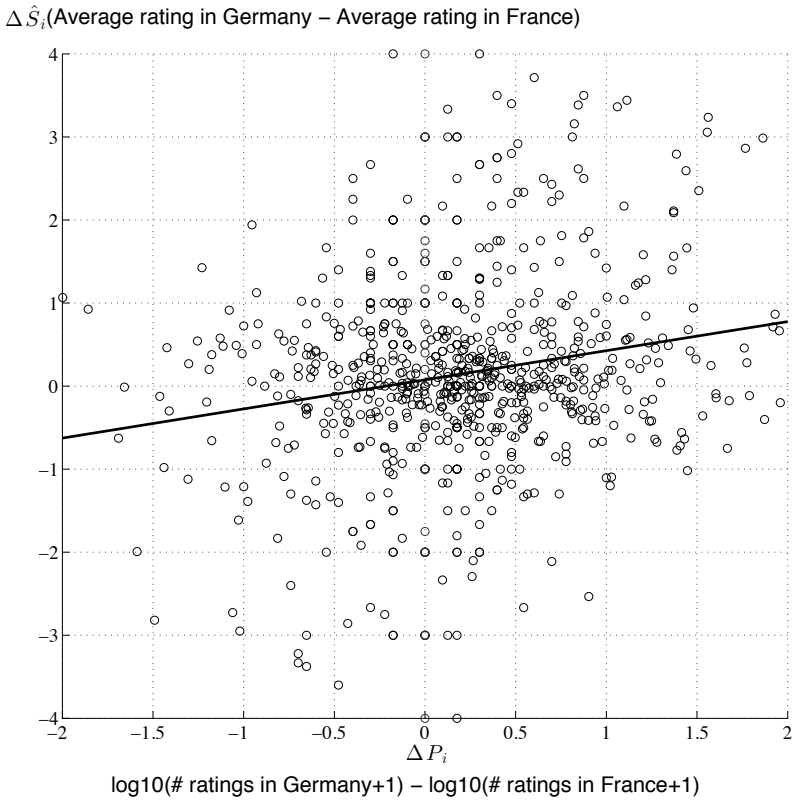

Fig. 5. **Scatterplot and regression line of the association between number of judgments and collective evaluation in the two-country data set.** Each circle corresponds to one product. For visual clarity, the graph is based on observations such that $-2 \leq \Delta P_i \leq 2$ ($N = 768$).

Discussion & Conclusion

When the crowd of judges forms endogenously, the accuracy of average ratings diminishes and systematic biases emerge. Our analyses focused on settings in which higher average ratings attract more additional ratings. This pattern of endogenous crowd formation is likely to prevail on widely used recommendation systems like review websites, as documented by our analyses of Amazon.com and Yelp.com data. Our model predicted that in such cases, a negative bias in average ratings would emerge and that this negative bias would affect particularly strongly items with few ratings. We also predicted that items with few ratings would suffer a ranking penalty. We found direct evidence of the pattern of evaluative biases predicted by our theory in an online experiment. We also found indirect evidence for our theory in analyses of field data.

The evaluative biases identified in this paper are consequential both for consumers of goods and services and producers because average ratings have been found to have a significant impact on purchase intensions (de Langhe, Fernbach, & Lichtenstein, 2016) and sales (Chevalier & Mayzlin, 2006; Dellarocas, Zhang, & Awad, 2007; Li & Hitt, 2008). When a superior item has received few ratings, it is likely to have a lower average rating than inferior items that received many ratings. Potential consumers who rely on average ratings to decide between choice alternatives will thus make systematic mistakes. The systematic bias we identified penalizes items with few ratings and favors items with many ratings. Overall, this contributes to a kind of rich-gets-richer dynamics that could lead items that are not necessarily of higher quality to obtain disproportionately high market shares (Salganik et al., 2006).

The evaluative biases we analyzed are produced by a mechanism similar to the 'hot-stove' effect, or the tendency of individuals to prematurely avoid alternatives with which they have obtained poor outcomes (Denrell, 2005; 2007; Denrell & March, 2001; Erev & Barron, 2005; Erev & Roth, 2014; Fiedler & Juslin, 2006; Hertwig, Barron, Weber, & Erev, 2004; Le Mens, Kareev, & Avrahami, 2016; March, 1996). Whereas existing work on the consequences of the hot-stove effect analyzed what happens to *individual* evaluations, here we analyzed what happens to *collective* evaluations when a *population* of evaluators prematurely avoids alternatives with poor ratings. This collective behavior leads to a negative

bias in average ratings similar to the negative bias in individual attitudes predicted by the hot stove effect. We conjecture that our findings are relevant not only to the dynamics of average ratings of products and services, but more generally to recommendation systems based on collaborative filtering algorithms, including those that control the display of articles on news websites or on social media platforms (Koren, Bell, & Volinsky, 2009).

The systematic nature of the evaluative biases we identified suggests that recommendation systems that report average ratings could improve their recommendations by adding a corrective term that compensates for the bias implied by endogenous crowd formation. For example, the graphs of Fig. 3 suggest that adding a positive 'bonus' to average ratings based on few ratings will improve the accuracy of collective evaluations. Another approach could be nudging users of online recommendation systems to write reviews about items that are likely to have been prematurely avoided (those with a low average rating based on few ratings) (Sunstein & Thaler, 2012). This would lower the endogenous character of crowd formation. The design and analysis of optimal corrections adapted to specific rating environments is an interesting avenue for future research.

References

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Blossfeld, H.-P., & Rohwer, G. (2001). Techniques of event history modeling. Mahwah, NJ: Lawrence Erlbaum Associates.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280. http://doi.org/10.1287/mnsc.2014.1909

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345–354.

de Langhe, B., Fernbach, P. M., & Lichtenstein, D. R. (2016). Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings. *Journal of Consumer Research*, *42*(6), 817–833. http://doi.org/10.1093/jcr/ucv047

Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23–45. http://doi.org/10.1002/dir.20087

Denrell, J. (2005). Why most people disapprove of me: experience sampling in impression formation. *Psychological Review*, *112*(4), 951–978. http://doi.org/10.1037/0033-295X.112.4.951

Denrell, J. (2007). Adaptive learning and risk taking. *Psychological Review*, *114*(1), 177–187. http://doi.org/10.1037/0033-295X.114.1.177

Denrell, J., & Le Mens, G. (2017). Information Sampling, Belief Synchronization, and Collective Illusions. *Management Science*, *63*(2), 528–547. http://doi.org/10.1287/mnsc.2015.2354

Denrell, J., & March, J. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*(5), 523–538. http://doi.org/10.1287/orsc.12.5.523.10092

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, *112*(4), 912–931. http://doi.org/10.1037/0033-295x.112.4.912

Erev, I., & Roth, A. E. (2014). Maximization, learning, and economic behavior. *Proceedings of the National Academy of Sciences*, *111*(Supplement_3), 10818–10825. http://doi.org/10.1073/pnas.1402846111

Fiedler, K., & Juslin, P. (2006). Taking the interface between mind and environment seriously. In K. Fiedler & P. Juslin (Eds.), (pp. 3–29). Information sampling and adaptive cognition.

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451.

Godes, D., & Silva, J. C. (2012). Sequential and temporal dynamics of online opinion. *Marketing Science*, *31*(3), 448–473. http://doi.org/10.1287/mksc.1110.0653

Hertwig, R., Barron, G., Weber, E., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539.

Kao, A. B., & Couzin, I. D. (2014). Decision accuracy in complex environments is often maximized by small group sizes. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1784), 20133305. http://doi.org/10.1098/rspb.2013.3305

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. *Computer*, *42*(8), 30–37. http://doi.org/10.1109/mc.2009.263

Le Mens, G., Kareev, Y., & Avrahami, J. (2016). The evaluative advantage of novel alternatives: an information-sampling account. *Psychological Science*, *27*(2), 161–168. http://doi.org/10.1177/0956797615615581

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product peviews. *Information Systems Research*, *19*(4), 456–474. http://doi.org/10.1287/isre.1070.0154

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025. http://doi.org/10.1073/pnas.1008636108/-/DCSupplemental

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. http://doi.org/10.1037/a0036677

March, J. (1996). Learning to be risk averse. *Psychological Review*, *103*(2), 309–319.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, *104*(8), 2421–2455.

McAuley, J., Pandey, R., & Leskovec, J. (2015a). Inferring Networks of Substitutable and Complementary Products (pp. 785–794). Presented at the the 21th ACM SIGKDD International Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/2783258.2783381

McAuley, J., Targett, C., Shi, Q., & van den Hengel, A. (2015b). Image-Based Recommendations on Styles and Substitutes (pp. 43–52). Presented at the the 38th International ACM SIGIR Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/2766462.2767755

Moe, W. W., & Schweidel, D. A. (2012). Online product opinions: incidence, evaluation, and evolution. *Marketing Science*, *31*(3), 372–386. http://doi.org/10.1287/mksc.1110.0662

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social Influence Bias: A Randomized Experiment. *Science*, *341*(6146), 647–651. http://doi.org/10.1126/science.1240466

Nielsen. (2013). *Global trust in advertising and brand messages* (pp. 1–16). New York: Nielsen.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854–856. http://doi.org/10.1126/science.1121066

Schlosser, A. E. (2005). Posting versus lurking: Communicating in a multiple audience context. *Journal of Consumer Research*, *32*(2), 260–265.

Sunstein, C. R., & Thaler, R. H. (2012). Nudge. Penguin UK.

Surowiecki, J. (2005). The Wisdom of Crowds. Anchor.

Acknowledgments

We are grateful to Julian McAuley for providing the Amazon.com data, to Shohei Yamamoto and Pablo López-Aguilar Beltran for assistance with programming the experiment; to participants at seminars at CMU, IESE, HEC Lausanne, U. Mannheim, MIT, NYU, SDU, UMass Amherst, Warwick, and at TOM 2016 conference for discussion and comments.

Author contributions

All authors contributed to the original idea and the design of the project. G. Le Mens, & B. Kovács contributed to the study design, and data analysis of the two analyses of field data. G. Le Mens performed the computer simulations and performed the study design, data collection and analysis of the experiments. G. Le Mens drafted the manuscript. J. Avrahami, Y. Kareev, and B. Kovács provided critical revisions. All authors approved the final version of the manuscript for submission.

Funding

G. Le Mens benefited from financial support from Spanish MINECO Grants PSI2013-41909-P and #AEI/FEDER UE-PSI2016-75353, a Ramon y Cajal Fellowship (RYC-2014-15035), and a Grant IN[15]_EFG_ECO_2281 from the Fundación BBVA. This paper was developed while G. Le Mens was on sabbatical at New York University, with support of grant CAS15/00225 from the Spanish Ministerio de Educacion, Cultura y Deporte. B. Kovács benefited from financial support from the Yale School of Management and Grant 100018_159511 from the Swiss National Science Foundation. Y. Kareev and J. Avrahami benefited from financial support from Israel Science Foundation Grant 121/11.

How endogenous crowd formation undermines the wisdom-of-the-crowd in online ratings

Gaël Le Mens [corresponding author], Balázs Kovács,
Judith Avrahami, Yaakov Kareev

SUPPLEMENTAL MATERIAL AVAILABLE ONLINE

COMPUTATIONAL ANALYSIS 1

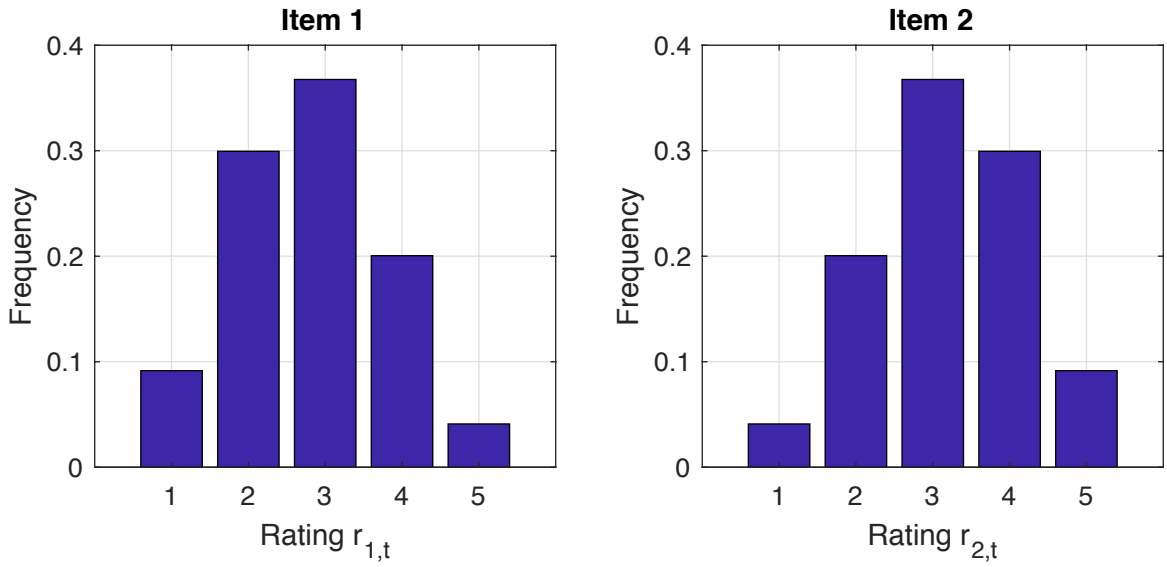

Fig. S1: Rating distributions for the two items in the simulations based on setting with two items and 10 judges. The true qualities are $q_1 = 2.8$ and $q_2 = 3.2$.

Base Model, $b = 0$.

This is a standard wisdom-of-the-crowd setup where the crowd is fixed at the outset (or at least forms in a way that does not depend on earlier ratings). The histograms in Figure S2A indicate that the numbers of judges for the two items (N_1 and N_2) have the same distribution. Moreover, the distribution of the difference in numbers of judges ($\Delta N = N_2 - N_1$) is symmetric around 0. The mean average ratings for the two items are equal to the true qualities ($E[\hat{S}_1] = 2.8$, $95\%CI = [2.80,2.80]$; $E[\hat{S}_1] = 3.2$, $95\%CI = [3.20,3.20]$). In other words, the average rating is an unbiased estimator of the true quality. The mean square error of the standard average rating is $MSE_{1,b=0} = 0.22$ for Item 1 and $MSE_{2,b=0} = 0.22$ for Item 2. This is much lower than the MSE of a single rating ($MSE_{1,single\ rating} = 0.99$; $MSE_{2,single\ rating} = 0.99$): judgment aggregation implies that average ratings tend to be closer to the true quality than individual ratings – the classic wisdom-of-the-crowd effect. The box plots in the top panel of Figure 1A shows that there is no systematic association between the number of judges for an item and the average rating obtained by the item: $corr(N_1, \hat{S}_1) = .003$ and $corr(N_2, \hat{S}_2) = .004$

Base Model, $b = -1$.

An opposite pattern of results emerges when judges are more likely to evaluate items with low average ratings ($b = -1$, Fig. S2B). In this case, average ratings systematically overestimate the qualities of the items ($E[\hat{S}_{1t}] = 2.90 > q_1 = 2.8$, $95CI = [2.90,2.90]$; $E[\hat{S}_{1t}] = 3.35 > q_2 = 3.2$, $95CI = [3.34,3.35]$). A negative association emerges between the number of judges and average ratings, as shown by the graphs of Fig. 1C: $corr(N_1, \hat{S}_1) = -.52$ and $corr(N_2, \hat{S}_2) = -.53$. Systematic mistakes emerge in a pattern that mirrors what was obtained with $b = 1$: when $N_2 \geq N_1 + 4$, $E[\hat{S}_2] < E[\hat{S}_1]$ and $P[\hat{S}_2 < \hat{S}_1] > P[\hat{S}_2 > \hat{S}_1]$. The mean square errors are larger than in the non-endogenous crowd formation case: $MSE_{1,b=-1} = 0.29$ and $MSE_{1,b=-1} = 0.37$. In this case as well, the wisdom-of-the-crowd effect is negatively impacted as compared to the non-endogenous crowd formation case.

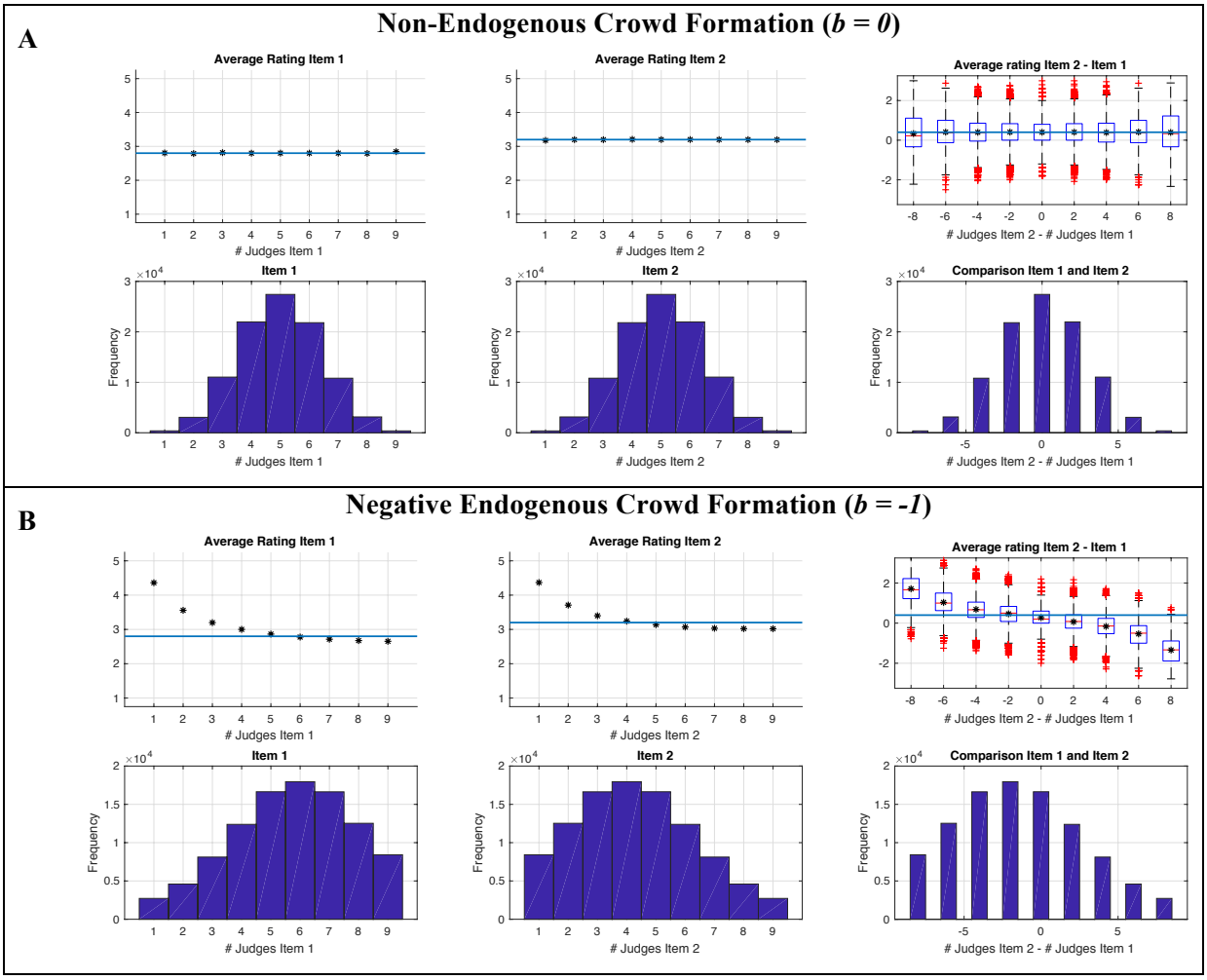

Fig. S2. Simulation of the model with 2 items. On the upper panel plots, the black stars indicate the means, and the horizontal lines indicates the true values (q_1, q_2 and $q_2 - q_1$ from left to right). On the boxplots, the boxes indicate the interquartile ranges.

Model with Social Influence Bias.

Let w be the direct social influence parameter. The probability of success of the binomial term, x_{it}, for Item i is:

$$(1 - w)p_i + w(\hat{S}_{ti} - 1)/4. \qquad \text{(S1)}$$

If $w = 0$, the model is the same as the model used in the above simulations. If $w = 1$, the mean of the distribution of the rating in period t, r_{it}, is the average rating $\hat{S}_{ti}$. If $0 < w < 1$ ratings are 'anchored' by the current average rating. We ran simulations with the same parameters as before and with $w = .5$. The results are reported in Fig. S3. We found that if $b = 0$, the social influence bias does *not* create any association between average rating and number of ratings. If $b > 0$, the social influence bias tends to further amplify the *positive* spurious association between number of ratings and average ratings created by our mechanism. If $b < 0$, the social influence bias further amplifies the *negative* spurious association between number of ratings and average ratings created by our mechanism. Overall, this analysis shows that our mechanism can operate in parallel with direct social influence. The social influence bias alone does not imply the emergence of an association between number of ratings and average ratings.

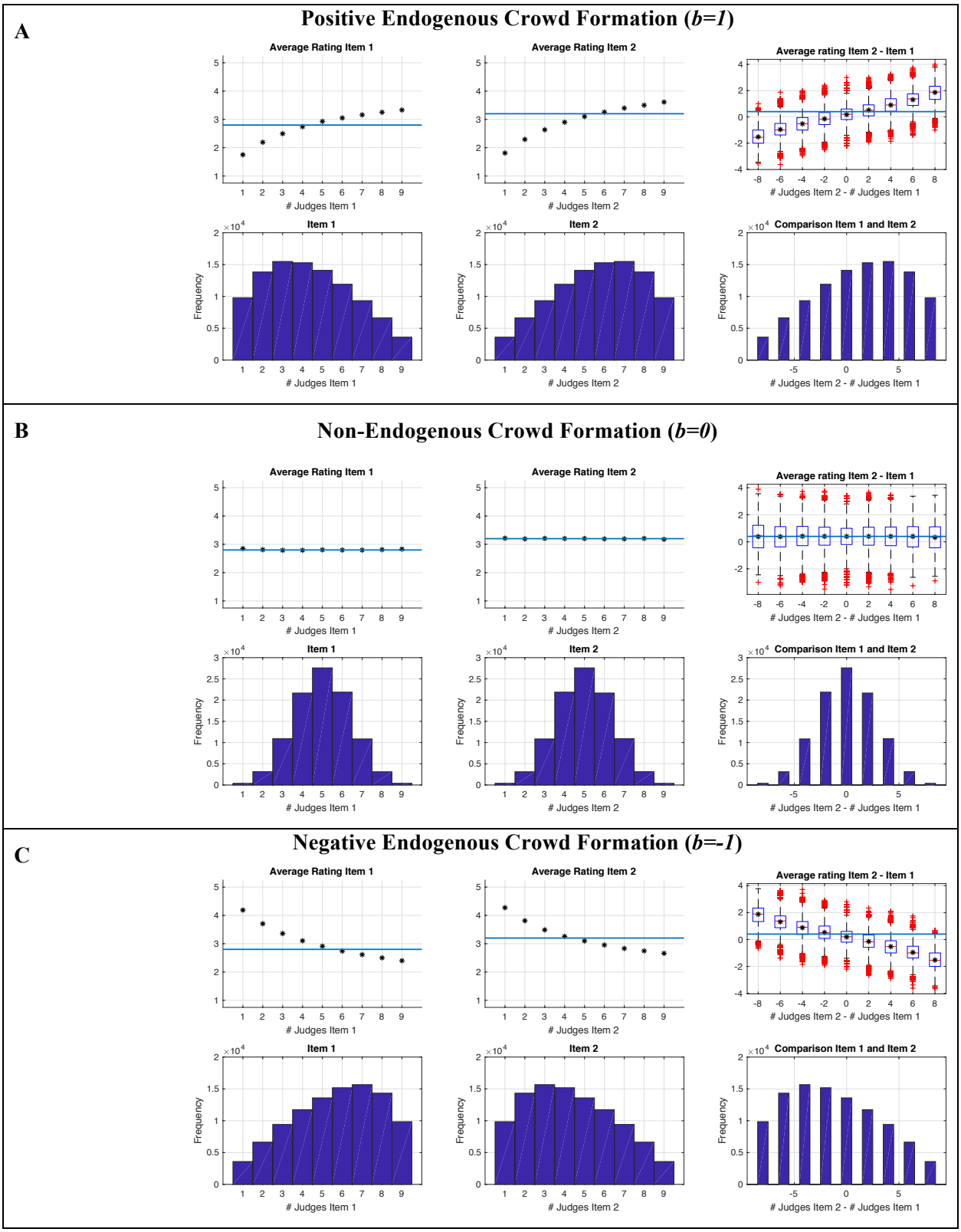

Fig. S3. Simulation of the model with 2 items and social influence bias ($w = 0.5$). On the upper panel plots, the black stars indicate the means, and the horizontal lines indicates the true values (q_1, q_2 and $q_{2-}q_1$ from left to right). On the boxplots, the boxes indicate the interquartile ranges.

Model with Direct Popularity Effect

Let d be the direct popularity effect parameter. The probability of success of the binomial term, x_{it}, for Item i is:

$$(1-d)p_i + d\frac{M-rank(\,N_{ti})}{M-1}, \qquad \text{(S2)}$$

where M is the number of available items, and $rank(\,N_{ti})$ is the rank of Item i, in terms of the number of ratings among all the available items. If there are just two items ($M = 2$), the probability of success of the item with the higher number of ratings is $(1-d)p_i + d(M-1)/(M-1) = p_i + d(1-p_i)$. The probability of success of the item with the lower number of ratings is $(1-d)p_i$. If $d = 0$, the model is the same as the model used in the above simulations. If $d = 1$, probability of success of the item is entirely driven by the rank of the item in terms of number of ratings. If $0 < d < 1$, the probability of success is affected both by the true quality (by construction, $p_i = (q_i - 1)/4$) and by the number of ratings received by the item N_{ti}. We ran simulations with the same parameters as before and with $d = .5$. The results are reported in Fig. S4. We found that if $b = 0$, the direct popularity effect creates a positive association between average rating and number of ratings. In other words, the direct popularity effect alone can explain the emergence of a positive association between number of ratings and average ratings. If $b > 0$, the direct popularity effect tends to further amplify the *positive* association between number of ratings and average ratings created by our mechanism. If $b < 0$, the direct popularity effect tends to attenuate the negative association between number of ratings and average ratings created by our mechanism. Because the direct popularity effect can produce, on its own, a spurious association between average rating and number of ratings similar to that produced by endogenous crowd formation, we designed our experiments such that this mechanism was unlikely to operate.

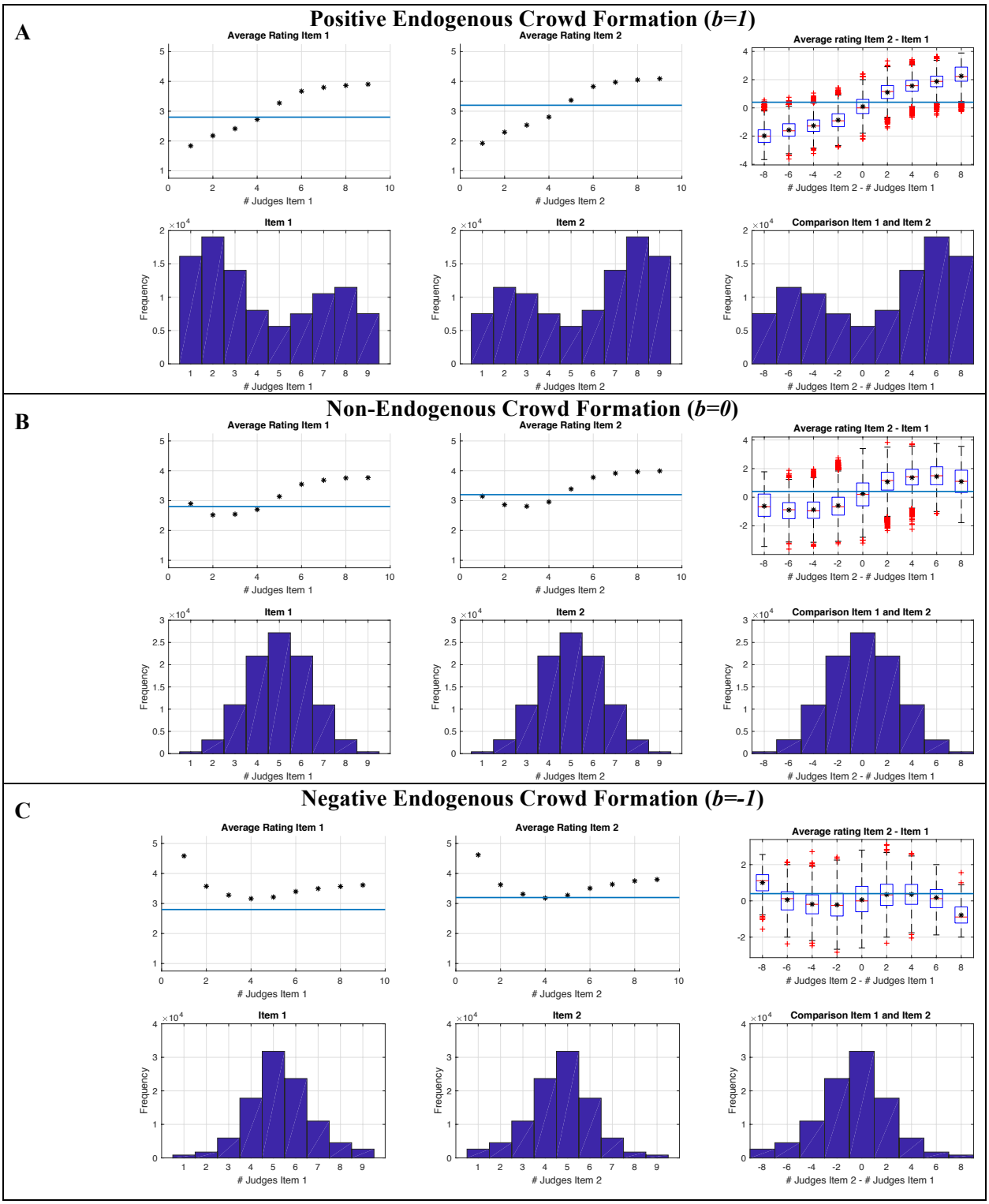

Fig. S4. Simulation of the model with 2 items and 'direct popularity effect' ($d = 0.5$). On the upper panel plots, the black stars indicate the means, and the horizontal lines indicates the true values (q_1, q_2 and $q_{2-}q_1$ from left to right). On the boxplots, the boxes indicate the interquartile ranges.

Base model with 100 judges

Consider the case where there are $N = 100$ judges, and $b = 1$ (all the other parameters remain the same). In this case, $corr(N_1, \hat{S}_1) = .47$ and $corr(N_2, \hat{S}_2) = .37$. This spurious association is not very consequential, however. The mean square errors are very small, ($MSE_{1,b=1} = 0.03$ and $MSE_{1,b=1} = 0.02$). This is because in almost all the simulation runs, the item with the fewer number of ratings has enough ratings for the average rating to be close to the true quality. The overall negative bias in average ratings is thus very small: ($E[\hat{S}_{1t}] = 2.78$, 95% $CI = [2.78,2.78]$; $E[\hat{S}_{1t}] = 3.19, 95\% CI = [3.19,3.19]$). And mistakes almost never occur: $P[\hat{S}_2 < \hat{S}_1] = .02$.

COMPUTATIONAL ANALYSIS 2

In the main body of the paper, we presented analyses of a simple model with two alternatives. On online recommendation systems, however, there are many alternatives. Here, using computer simulations, we show that even with more than two alternatives, positive endogenous crowd formation implies that items with comparatively few ratings are subject to a stronger negative bias in average ratings than items with many ratings. Moreover, systematic ranking mistakes emerge: items with comparatively few ratings suffer from a ranking penalty.

The task environment of the simulations is very similar to the task environment of the experiment reported in body of the paper. Simulation results will be used to formulate empirical predictions tested in the experiment.

Simulation Model

The model is the same as that in the main body of the paper, except for the number of items and judges. Suppose there are 50 items and 500 judges. The qualities of the items follow a $Beta(8,5)$ distribution rescaled between 1 and 5. We use this distribution because it is similar to the quality distribution of the items used in the experiment reported in the paper. As before, the judges join the system sequentially, one in each period. Each judge rates one item at the time he or she joins. The probability that Item i is selected is proportional to

$$\frac{e^{b\,\hat{S}_{it}}}{\sum_{j=1}^{50} e^{b\,\hat{S}_{jt}}}.$$

All the other components of the model remain the same as before.

We analyze three scenarios, with $b = 1$ (higher average ratings attract additional ratings: positive endogenous crowd formation), $b = -1$ (lower average ratings attract additional ratings: negative endogenous crowd formation) and $b = 0$ (the flow of ratings is independent form the average rating: non-endogenous crowd formation). We ran 2,000 simulations of the model in each scenario. Simulation results are summarized in Table 1.

Overall bias in average rating

Let $\hat{S}_i$ denote the final average rating for item *i* and ε_i^S denote the error in average rating: $\varepsilon_i^S = \hat{S}_i - q_i$.

When there is positive endogenous crowd formation ($b = 1$), there is a systematic negative bias in average rating: the mean error is negative ($E[\varepsilon_i^S] = -0.15$). By contrast, when there is negative endogenous crowd formation ($b = -1$), there is a positive bias in average ratings: the mean error is positive ($E[\varepsilon_i^S] =$

0.12). Finally, when there is non-endogenous crowd formation ($b = 0$), there is no bias in average ratings ($E[\varepsilon_i^S] = 0.0007$).

Table S1. Simulation Results.

		Non-endogenous crowd formation process	Endogenous crowd formation process	
			Positive (items with higher average rating receive more additional ratings)	Negative (items with higher average rating receive fewer additional ratings)
		$b = 0$	$b = 1$	$b = -1$
Errors in average rating $\varepsilon_i^S = \hat{S}_i - \hat{q}_i$	$E[\varepsilon_i^S]$	-0.0007 [-0.003, 0.001]	-0.15 [-0.15, -0.14]	0.12 [0.12, 0.13]
Frequencies: Underestimation / Overestimation ($N = 50$)		25/25	28/22	21/29
Association between log10 of the number of ratings and average rating $\varepsilon_i^S = \beta\, log_{10}(1 + N_i) + \gamma + \epsilon_i$,	$E[\beta]$	0.0001 [-0.02, 0.02]	1.05 [1.04, 1.06]	-0.85 [-0.86, -0.84]
Association between ranking error ($\varepsilon_i^{rank} = rank(S_i) - rank(\hat{q}_i)$) and rank in number of ratings $rank(N_i)$: $\varepsilon_i^r = \beta^r\, rank(N_i) + \gamma^r + \epsilon_i$	$E[\beta^r]$	0.001 [-0.002, 0.007]	0.17 [0.17, 0.17]	-0.18 [-0.19, -0.18]

Note: 95% CI in brackets.

Systematic pattern of evaluative biases:

Consider the outcome of just one simulation run. We regress the error in average ratings ε_i^S on the logarithm of the number of ratings, $log_{10}(1 + N_i)$ (we take the log because the distribution of the number of ratings is skewed) and quality. There are 50 items. We estimate the following equation:

$$\varepsilon_i^S = \beta\, log_{10}(1 + N_i) + \gamma + \epsilon_i, \quad \text{(S3)}$$

where β and γ are the coefficients to be estimated and ϵ_i is an error term. The estimated β coefficient is a measure of the strength of the association between number of ratings and the error in average ratings.

When there is positive endogenous crowd formation ($b = 1$), a positive association between number of ratings and average ratings emerges ($E[\beta] = 1.05$). The association is positive in all simulation runs. In other words, the negativity bias is stronger for items with comparatively few ratings. The pattern is opposite when there is negative endogenous crowd formation ($b = -1$). In this case a negative association between number of ratings and average ratings emerges ($E[\beta] = -0.85$). The association is negative in all simulation runs. (The effect with *b = -1* is not the exact opposite than what was obtained with *b = 1*, because the quality distribution is skewed – ancillary simulations show that if the quality

distribution is symmetric the effects obtained for $b = 1$ and $b = -1$ are exact opposites). Finally, when there is non-endogenous crowd formation ($b = 0$), there is no bias in average ratings ($E[\beta] = 0.0001$).

Systematic pattern of ranking mistakes.

Let $rank(S_i)$ be the rank of item *i* in terms of average ratings. If there are 50 items, $rank(S_i) = 50$ for the item with highest average rating and $rank(S_i) = 1$ for the item with the lowest average rating. We define the error in rank as follow $\varepsilon_i^r = rank(S_i) - rank(\hat{q}_i)$. Therefore, if $\varepsilon_i^{rank} > 0$, the rank of item *i*, in terms of average rating, is better than its true rank. Conversely, if $\varepsilon_i^{rank} < 0$, the rank of item *i*, in terms of average rating, is worse than its true rank. For each simulation run, we estimate the following equation

$$\varepsilon_i^r = \beta^r \, rank(N_i) + \gamma^r + \epsilon_i, \qquad \text{(S4)}$$

where β^r and γ^r are the coefficients to be estimated and ϵ_i is an error term. The estimated β^r coefficient is a measure of the strength of the association between ranking error and the rank in number of ratings.

When there is positive endogenous crowd formation ($b = 1$), a positive association between number of ratings the ranking error emerges ($E[\beta^r] = 0.17$): items with few ratings are likely to have a rank that is worse that their quality ranking. Items that were in the smaller half of the distribution of the number of ratings had an average rating rank worse than their quality rank (the mean penalty across items and simulation runs was $E\left[\varepsilon_i^{rank}\right] = -2.0$). By contrast, items that were in the larger half of the distribution of the number of ratings had an average rating rank better than their quality rank (the mean advantage across items and simulation runs was $E\left[\varepsilon_i^{rank}\right] = 2.0$).

When there is negative endogenous crowd formation ($b = -1$), the pattern is opposite. In this case a negative association between number of ratings the ranking error emerges ($E[\beta^r] = -0.18$). Items with few ratings are likely to have a rank that is better that their quality ranking.

When crowd formation is not endogenous ($b = 0$), no systematic association between number of ratings the ranking error emerges ($E[\beta^r] = 0.001$).

Direct Popularity Effect

We saw in the previous section that the 'direct popularity effect' could create an association between number of ratings and average ratings across simulation runs. This suggests that this mechanism too could create a spurious association, across items, between number of ratings and average ratings. To check this prediction, we ran 1,000 simulations of the model with $b = 0$ (non-endogenous crowd formation) and $d = 0.5$ (positive direct popularity effect – see eq. S2). We estimated eq. S3 on data produced by all the simulation runs. We found all 1,000 estimated beta coefficients to be positive. This indicates that the 'direct

popularity effect' can also produce a spurious association between number of ratings and average ratings across items, like endogenous crowd formation. This implies that finding empirical evidence of a spurious association between number of ratings and average ratings across items does not unambiguously indicate endogenous crowd formation. Therefore, we designed our experiment such that the 'direct popularity effect' was unlikely to operate.

EXPERIMENT

Methodological Details

We recruited 454 participants on Amazon Mechanical Turk for this experiment. It was administered via a Qualtrics survey embedded in the Amazon Mechanical Turk webpage as an iframe. After signing up for the task, participants read the informed-consent form. Then they were randomly assigned to one of 9 conditions. Conditions 1 to 4 consisted in "worlds" in which participants selected which picture to rate (as described in the body of the paper). The 5 remaining conditions were the 'independent' conditions in which participants could not choose which picture to rate (they had to rate 10 pictures shown sequentially to them).

The 4 Worlds:

After clicking a button (see the screenshot in Fig. S5), participants were shown the corresponding picture and gave it a rating of 1 to 5 stars (see the screenshot in Fig. S6). Participants rated 10 pictures one after the other. Finally, they answered a few biographical questions. When a participant completed the survey, his or her ratings were sent to a webserver, where the average rating for each of the 10 pictures was updated. When the following participant logged in to the survey, the average ratings for the 50 pictures were read from the server and used to order the buttons in the experimental display. On average, participants completed the survey in 267 s (σ = 96 s).

The 5 independent conditions:

Participants rated 10 pictures one after the other. Pictures were shown in a random sequence. Participants were not provided with any information about the behavior of other participants in the experiment. On average, participants completed the survey in 233 s (σ = 92 s).

Fig. S5.
Experiment: Example of the main screen (selection of which picture to rate).

Fig. S6.
Experiment: Example of a picture rating screen.

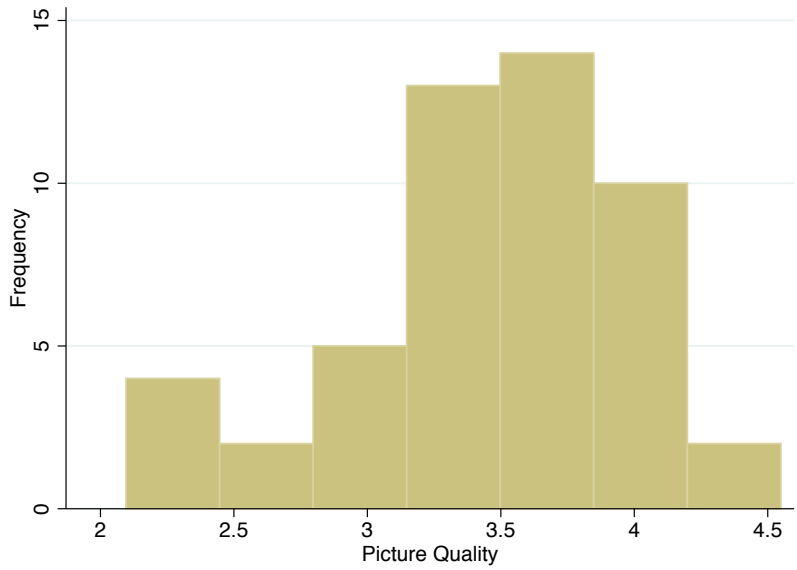

Fig. S7.
Experiment: Histogram of the picture quality estimates constructed from the ratings obtained in conditions 5_1 to 5_5 (N=50).

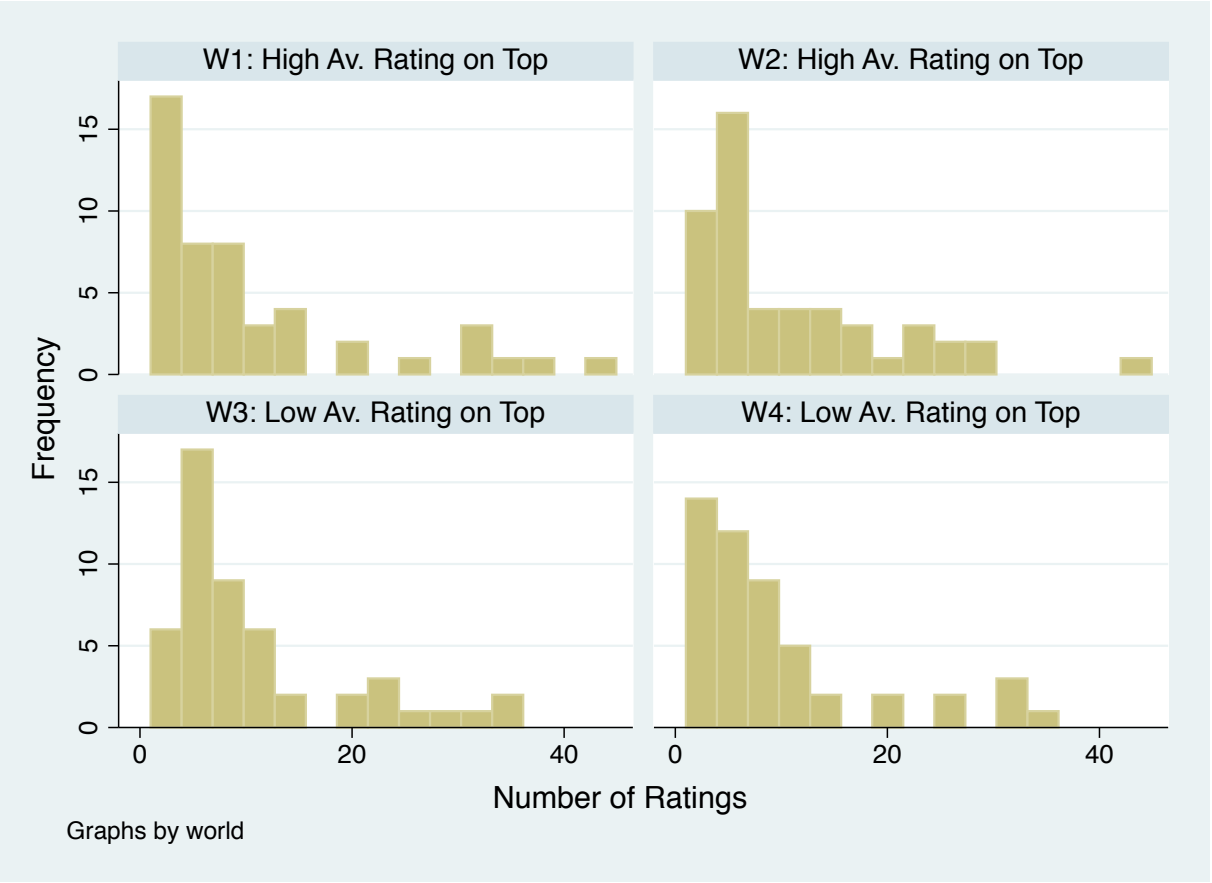

Fig. S8.
Experiment: Histograms of the final number of ratings (N=50,50,50,50).

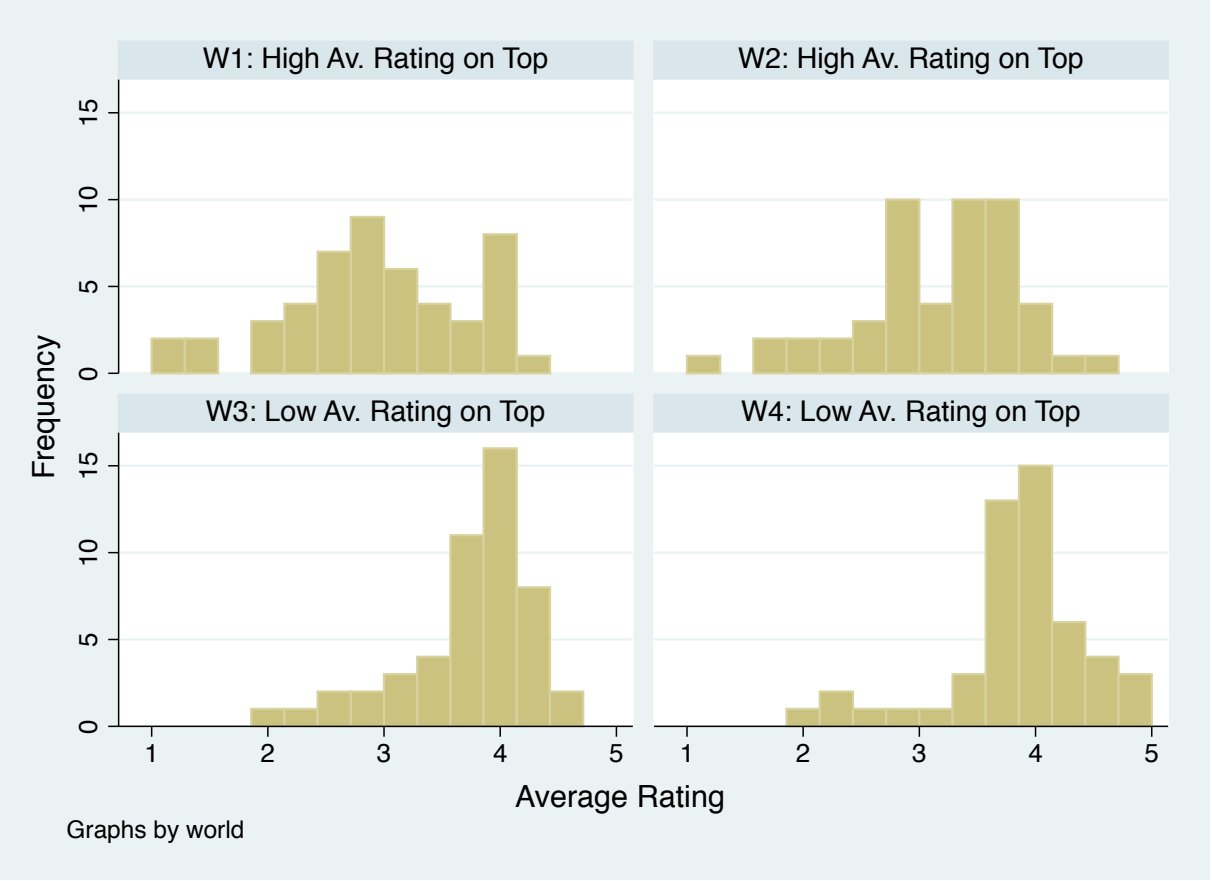

Fig. S9.
Experiment: Histograms of the final average rating (N=49,50,50,50).

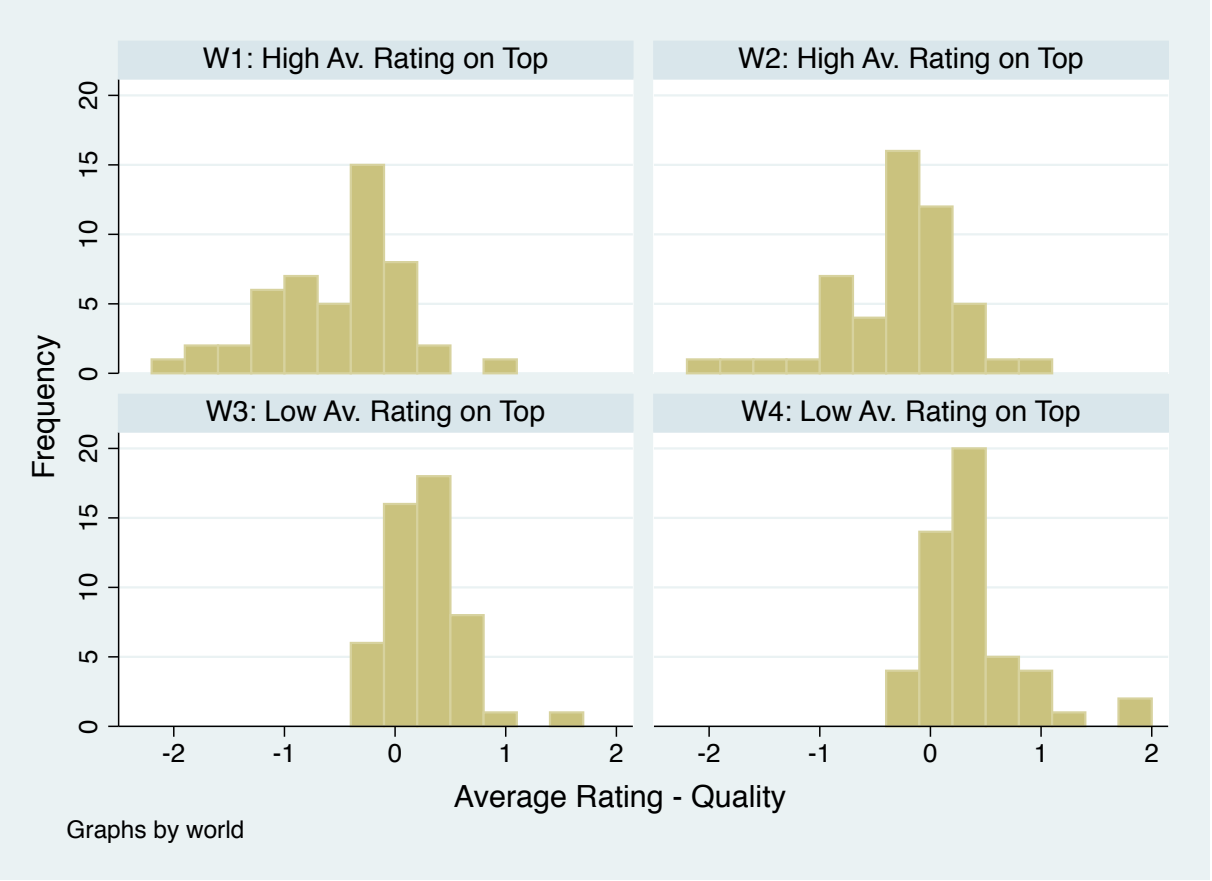

Fig. S10.
Experiment: Histograms of the error in average rating (N=49,50,50,50).

Spurious Association between Number of Ratings and Average Rating

In the section 'Analysis of Field Data 2' in the body of the paper, we report analyses that measure the spurious association between number of ratings and average rating using online ratings about the same products, but from two different websites. Here, we illustrate how this approach works on our experimental data. This provides a proof-of-concept for the use of the method on field data.

First, consider the two worlds with positive endogenous crowd formation (Worlds 1&2). Our theory predicts the emergence of a positive spurious association between number of ratings and average ratings. We can measure it by regressing the difference in average rating on the difference in number of ratings (see eq. 7 in the body of the paper). Parameter estimates indicate a positive association ($\beta = 1.05, 95\%\ CI = [.72, 1.39]$, see left panel of Fig. S11). The result is opposite when there is negative endogenous crowd formation (Worlds 3&4). The spurious association is negative ($\beta = -0.74, 95\%\ CI = [-1.04, -0.44]$, see right panel of Fig. S11).

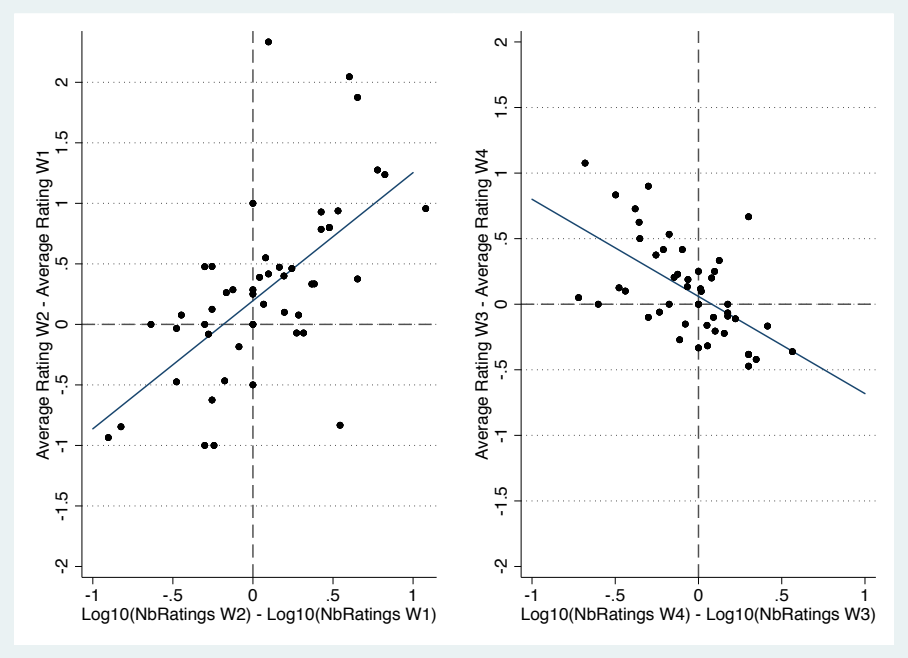

Fig. S11. Experiment: Scatterplots of them spurious association between number of ratings and average ratings in pairs of worlds with endogenous crowd formation. Left Panel: Worlds 1&2. Right Panel: Worlds 3&4. Each circle corresponds to one of the 50 pictures. The blue line is the best-fitting regression line.

ANALYSIS OF FIELD DATA 1

Rating data from Amazon.com

This data set was collected by Julian McAuley by means of a web crawler that generated product ID numbers (ASIN numbers) and downloaded all the published reviews for the identified products (18,19). It is publicly available at http://jmcauley.ucsd.edu/data/amazon/. We retrieved it on Feb. 16, 2016.

We eliminated all observations for which product ID, time of posting online**,** or category information was not available. We also removed all observations in the category labeled "Buy a Kindle" and all observations with the ASIN number B00EOE0WKQ (corresponding to the "Amazon Fire Phone") because inspection of the data revealed that they do not correspond to actual product ratings (these "observations" were created by a bug in the program written to download data from Amazon.com). The final data set included 77,968,408 ratings of 8,498,036 products in 82 product categories by 20,556,716 individual reviewers (i.e., different user IDs). Examples of product categories include "Electronics," "Books," and "Video Games." We included dummy variables for each category in our estimations based on the whole data set to control for differences in dynamics across categories (see table S3 for a list of the categories with at least 100,000 ratings).

Rating Data from Yelp.com

This data set contains 2.2 million ratings of local businesses made available to the research community by Yelp.com. It was released to the research community for the Yelp 7th Dataset Challenge. The data set is available at https://www.yelp.com/dataset_challenge. We retrieved it on Feb. 16, 2016.

Examples of business categories in this data set include "Bars," "Auto Repair," and "Health Markets." The final data set included 2,225,213 ratings of 77,079 businesses in 488 business categories by 552,339 individual reviewers (i.e., different user IDs)

Additional Comments on the Results

Although we cannot provide irrefutable evidence that in these data, the average rating has a causal effect on the arrival of new ratings, existing research suggests that potential reviewers are likely to review items with high average ratings even after accounting for heterogeneity in quality (Moe & Schweidel, 2012). Besides, reviewed items are experience goods (e.g., restaurant dinners, books) in the sense that quality is not observable before experiencing the item. Amazon and Yelp are among the most widely consulted sources of information about the experiences of other users. This suggests that there is a causal effect of average ratings on the flow of ratings.

Third, ancillary linear regressions of the waiting time between two rating instances on the valence of collective evaluation show a negative effect of the valence of collective evaluations on the waiting time, even with product fixed effects. This pattern is consistent with the findings reported in the body of the paper. We consider these analyses as 'ancillary' because such estimations are subject to right-censoring (the waiting time between the last rating and the date at which we collected the data is unspecified). We still mention them because we can include product fixed effects in such estimations. By contrast, it is not possible to include fixed effects in estimations of the hazard rate models, because those rely on maximum likelihood estimations which are computationally intractable if we included millions of fixed effect parameters. Overall, both types of estimations (hazard rate models without fixed effects and linear regressions of waiting times with fixed effects) provide results consistent with the premise of the theory that average ratings tend to attract more additional ratings.

Table S2.

Parameter estimates from hazard rate models predicting the occurrence of rating instances according to eq. 1 in the Amazon.com and Yelp.com data. Note that the body of the article reports the effects in terms of hazard ratios, whereas the table reports estimates of the coefficients. (e.g., for the Amazon.com data, a 1-point increase in average rating implies a rating flow that is equal to the baseline multiplied by exp(0.15), that is, a flow of 1.16, or a 16% increase.

	Amazon.com	Yelp.com
	Model 1	Model 2
Average rating$_{it}$	0.15 (0.0015)	0.13 (0.0045)
Log_{10} (# of ratings$_{it}$)	2.53 (0.0029)	2.24 (0.0075)
Number of items	8,498,036	77,079
Number of unique reviewers	20,652,778	552,339
Number of ratings	77,968,406	2,225,213
Number of item categories	82	488
Fixed effects	Yes	Yes

Note: Standard errors (clustered by item) in parentheses.

Table S3.

Parameter estimates from hazard rate models predicting the occurrence of rating instances in the Amazon.com data set. Each row presents the estimated effect of average rating (β) from a model that was like Model 1 in table S2 but included only the product category listed in the first column**.** Results are shown only for those categories with at least 100,000 ratings.

CategoryName	Effect of Average Rating_{it}	Number of items	Number of ratings
Amazon Instant Video	0.22 (0.021)	23 909	580 351
Appliances	0.21 (0.0096)	11 369	142 971
Apps for Android	0.37 (0.015)	61 202	2 632 219
Arts Crafts & Sewing	0.096 (0.0052)	112 034	507 506
Automotive	0.16 (0.0031)	319 125	1 367 272
Baby	0.17 (0.0078)	64 366	913 581
Baby Products	0.16 (0.029)	9 350	152 860
Beauty	0.14 (0.0041)	248 745	2 015 916
Books	0.15 (0.0032)	2 328 220	22 473 664
CDs & Vinyl	0.17 (0.0044)	485 138	3 740 856
Cell Phones & Accessories	0.17 (0.0059)	318 824	3 426 881
Clothing Shoes & Jewelry	0.096 (0.0019)	1 081 518	5 353 952
Digital Music	0.067 (0.0067)	256 605	530 381
Electronics	0.19 (0.0046)	472 843	7 608 785
Grocery & Gourmet Food	0.19 (0.0061)	165 826	1 293 866
Health & Personal Care	0.20 (0.0047)	251 167	2 962 600
Home & Kitchen	0.19 (0.0039)	409 675	4 242 649
Industrial & Scientific	0.15 (0.0078)	45 157	261 110
Movies & TV	0.18 (0.0086)	187 550	4 433 587
Musical Instruments	0.17 (0.0068)	66 752	436 581
Office Products	0.21 (0.0055)	129 258	1 230 958
Patio Lawn & Garden	0.17 (0.0047)	105 549	984 668
Pet Supplies	0.16 (0.0062)	102 996	1 231 641
Software	0.11 (0.018)	17 432	282 406
Sports & Outdoors	0.14 (0.0029)	475 978	3 231 290
Tools & Home Improvement	0.15 (0.0032)	259 501	1 914 488
Toys & Games	0.11 (0.0044)	326 700	2 242 872
Video Games	0.14 (0.013)	48 116	1 179 508

Note: Standard Errors (clustered by product) in parentheses. Estimations include dummy variables for calendar year and product age in years.

Table S4.

Parameter estimates from hazard rate models predicting the occurrence of rating instances in the Yelp.com data set. Each row presents the estimated effect of average rating (β) from a model that was like Model 2 in table S2 but included only the product category listed in the first column**.** Results are shown only for those categories with at least 10,000 ratings.

CategoryName	Effect of Average Rating$_{it}$	Number of items	Number of ratings
Active Life	-0.0050 (0.015)	2,801	50,012
American (New)	0.44*** (0.053)	619	50,551
American (Traditional)	0.21*** (0.028)	896	49,555
Arts & Entertainment	-0.042 (0.026)	1,491	80,938
Asian Fusion	0.40*** (0.065)	254	17,693
Auto Repair	-0.019 (0.012)	1,492	22,639
Automotive	0.063*** (0.017)	1,348	14,607
Bakeries	0.17*** (0.030)	888	31,729
Barbeque	0.41*** (0.074)	264	14,683
Bars	0.19*** (0.021)	2,542	156,336
Beauty & Spas	0.19*** (0.021)	1,318	26,371
Breakfast & Brunch	0.34*** (0.035)	924	86,336
Buffets	0.23** (0.078)	240	27,618
Burgers	0.23*** (0.027)	1,580	77,986
Chinese	0.25*** (0.032)	1,003	31,714
Delis	0.24*** (0.048)	436	14,445
Doctors	0.0064 (0.012)	1,326	11,303
Fast Food	0.11*** (0.019)	1,712	24,040
Food	0.14*** (0.014)	7,618	212,401
Greek	0.40*** (0.048)	373	17,891
Hair Salons	0.12*** (0.017)	1,986	29,028
Health & Medical	-0.032** (0.012)	1,568	13,938
Home Services	0.039*** (0.011)	1,588	14,030
Hotels & Travel	-0.067*** (0.017)	2,223	89,861
Indian	0.49*** (0.069)	285	12,604
Italian	0.33*** (0.034)	525	33,212
Japanese	0.30*** (0.058)	275	13,287
Korean	0.39*** (0.077)	200	15,598
Local Services	0.063*** (0.013)	3,167	31,648
Mexican	0.22*** (0.024)	1,732	88,690
Nightlife	0.34*** (0.056)	462	29,267
Pizza	0.21*** (0.021)	1,426	43,401
Pubs	0.20*** (0.048)	633	23,673
Restaurants	0.22*** (0.029)	1,315	53,776
Sandwiches	0.25*** (0.046)	678	23,375
Seafood	0.30*** (0.039)	513	47,168
Shopping	0.043*** (0.012)	3,295	33,921
Skin Care	0.088** (0.032)	574	10,783
Steakhouses	0.27*** (0.038)	340	38,587
Sushi Bars	0.35*** (0.044)	582	50,429
Thai	0.45*** (0.059)	485	36,523
Vegetarian	0.24*** (0.065)	177	18,452
Vietnamese	0.41*** (0.069)	280	19,514
Wine Bars	0.32*** (0.088)	162	12,637

*Note: Standard Errors (clustered by business) in parentheses. Estimations include dummy variables for calendar year and business age in years. *: p<.05, ** : p<.01, ***: p<.001.*

ANALYSIS OF FIELD DATA 2

Data:

We obtained product ratings from two Amazon websites: Amazon.de (Germany) and Amazon.fr (France). Products listed on these two websites share the same ASIN numbers. This makes it easy to match products across the sites. We decided to examine the ratings of iPad cases because we believed that would-be Amazon consumers are likely to consult the reviews before purchasing one of these items. This conjecture was based on several observations: There are numerous models of iPad cases, at roughly the same price (around 20 euros), and they tend to look alike. An iPad case is an item people rely on to protect a fairly expensive electronic device, so there are good reasons for wanting a truly protective case. Thus, there is value in relying on the experience of other consumers before deciding which model to purchase. Finally, Amazon lists more product reviews about this kind of item than almost any other website and thus likely attracts many would-be consumers who want to get some information before making their decision.

Another reason for choosing this product category was that iPad cases are commodity items and are highly similar substitutes to each other, and thus it is likely that there is little cultural difference between countries in taste for one case or another. This intuition was confirmed by the fact that the correlation between the average rating on Amazon.de and the average rating on Amazon.fr was very high for items that had received many ratings in both countries (Fig. S12).

We first compiled a list of all iPad cases by searching for the term "iPad case" within the category "Computers & Accessories : Accessories : Tablet Accessories : Bags, Cases & Sleeves : Cases" on Amazon.co.uk. This search yielded 4823 product identifiers referring to 1036 unique products. The reason there were fewer unique products than product identifiers is that Amazon gives separate identifiers (ASIN numbers) to different versions (e.g., colors or shapes) of the same iPad case but treats them as the same product when it comes to showing ratings and reviews on the product page—the ratings of the different versions of a case are aggregated into a unique average rating that is shown on the pages of all the versions of that product. For example, the average rating of a red version of a given model is derived from all the ratings posted for that model in all the available colors. And the average rating of that red iPad case is exactly the same as the average rating of the blue version of the same model. To eliminate duplicates, we examined the text of the first review each product had received. If several items with different identifiers had the same number of ratings, the same average rating, and the same text in the first review, we kept only one of those observations. We then downloaded the ratings for all of these unique ASINs from the German (Amazon.de) and the French (Amazon.fr) sites.

We found 788 unique products that had at least one rating on both Amazon.de and Amazon.fr at the date on which we downloaded the data (May 23, 2016). In total, this procedure yielded 76,195 reviews.

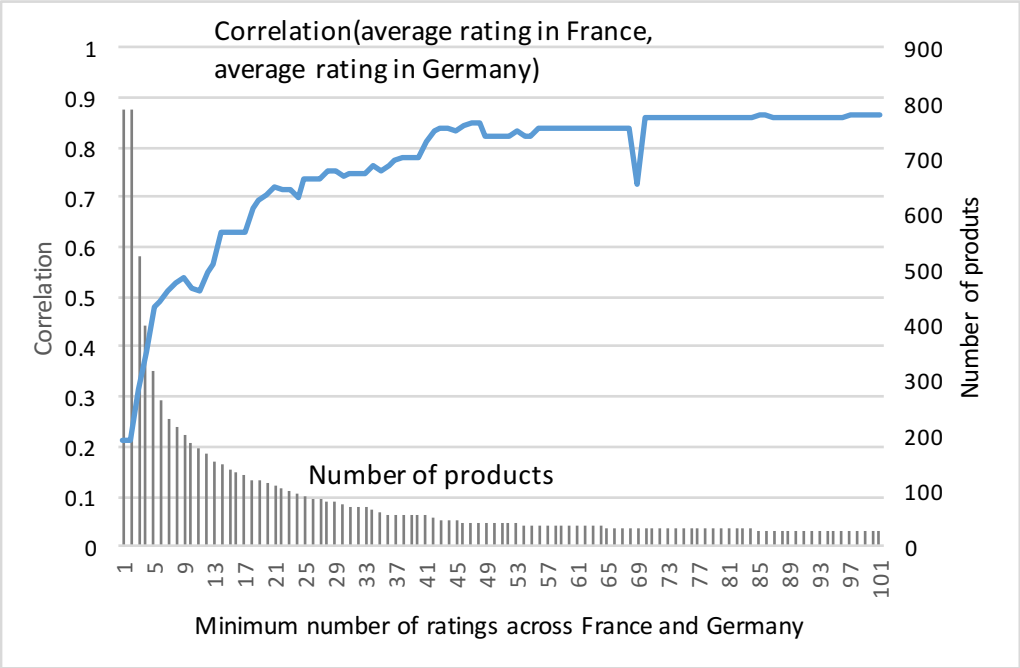

Fig. S12.
Correlation between the average ratings on Amazon.de and Amazon.fr. Each correlation is based on only those products for which both countries had at least the number of ratings indicated on the *x*-axis (the number of such items is indicated by the height of the corresponding vertical grey bar).

Methodological Details: Analysis of the Effect of the Average Rating on the Arrival of New Ratings.

Model estimations are reported in table S5. Model 1 included only the average rating and a country dummy. Estimation of this model indicated that there was a strong positive association between average rating and the arrival rate of new judgments.

We also estimated a model in which we included popularity ($\log_{10}$ of 1 + the number of ratings) as a predictor. In this case as well, we found a positive coefficient for average rating (Model 2, table S5). Although this suggested that the displayed average rating had a causal effect on the flow of ratings beyond the effect of popularity information, this latter result is not crucial to our prediction. The only relation that is needed for our prediction regarding the emergence of a negative evaluative bias is that there is a *probabilistic* dependence between occurrence of rating instances and average rating, *conditional on quality*. This was established by the estimation of Model 1.

Table S5.

Parameter estimates from conditional logit regressions predicting the country of a rating instance in the two-country data set.

	Model 1	Model 2
Average rating$_{it}$	0.64*	0.25***
	[0.057,1.23]	[0.17,0.34]
Log_{10} (1+# of ratings$_{it}$)		2.12***
		[1.97,2.27]
Country (=1 if France)	-0.66***	0.016
	[-0.99,-0.34]	[-0.063,0.095]
Log-likelihood	-36,749	-26,173
Pseudo R^2	0.13	0.38
Number of items	700	700
Number of unique reviewers	56,496	56,496
Number of ratings	60,670	60,670

*Note: 95CI in brackets. *: p<.05, ** : p<.01, ***: p<.001.*

Additional Analyses: Association between Average Rating and Number of Ratings Conditional on Quality

Here, we complement the linear regression of $\Delta\hat{S}_{i,t}$ (difference in collective evaluations) on $\Delta P_{i,t}$ (difference in popularities) reported in the body of the article (Fig. 3) with the results of linear regressions of average rating on popularity with product fixed effects (see table S6). The product fixed effects allowed us to control for the effect of unobserved quality. In the simplest case (Model 1), the fixed-effect linear regression is equivalent to the linear regression of $\Delta\hat{S}_{i,t}$ on $\Delta P_{i,t}$ displayed on Fig. 3. Formally, we estimated the following equation based on data at the end of the observation window:

$$\hat{S}_{ic} = \beta P_{ic} + \pi_i + \omega_c + \epsilon_{ic}, \quad \text{(S5)}$$

where $\hat{S}_{ic}$ is the average rating of product *i*, in country *c*, P_{ic} is the popularity of product *i* in country *c* ($\log_{10}$ of 1 + the number of ratings in country *i*), π_i is a product fixed effect, ω_c is a country fixed effect, and ϵ_{ic} is the error term**.** In Model 2, we included a quadratic term ($P_{ic}{}^2$) and obtained a negative estimated coefficient. This indicated that the association between popularity and collective evaluation was of decreasing strength as popularity became larger, consistent with the pattern of Fig. 1B. Model 3 was an estimation of the basic equation (i.e., Model 1) based on items that had at most 10 ratings in the focal country, whereas Model 4 was the same regression performed on items that hade more than 10 ratings in the focal country. Model 3 showed that the association between popularity and collective evaluation was strongly positive and significant for items with low levels of popularity (β = 0.81, 95CI = [0.34,1.28], N = 354). Model 4 showed that there was essentially no association when the number of ratings was higher than 10 (β = 0.051, 95CI = [-0.075,0.18], N = 166).

Table S6.

Parameter estimates from linear regressions of average rating on popularity (1 + Log_{10} of the number of ratings) at the time the data was downloaded from Amazon.de and Amazon.fr. The fixed effects are for each product. Estimations of Models 1 and 2 are based on all the products**.** The estimation of Model 3 was based on the products with at most 10 ratings in the focal country. The estimation of Model 4 was based the products with more than 10 ratings in the focal country.

	Model 1	Model 2	Model 3	Model 4
Log_{10} (1+# of ratings$_{it}$)	0.35***	0.80***	0.81***	0.051
	[0.23,0.47]	[0.47,1.13]	[0.34,1.28]	[-0.075,0.18]
Log_{10} (1+# of ratings$_{it}$)^2		-0.18**		
		[-0.30,-0.055]		
Country (=1 if France)	-0.075	-0.079	-0.045	-0.053
	[-0.16,0.010]	[-0.16,0.0067]	[-0.20,0.11]	[-0.12,0.013]
Log-likelihood	-1,394.70	-1,386.70	-778.00	68.30
Number of items	788	788	354	166
Product Fixed Effects	Yes	Yes	Yes	Yes

*Note: 95CI in brackets. *: p<.05, ** : p<.01, ***: p<.001.*

Table S7.

Parameter estimates from linear regressions of rating values on popularity (1 + Log_{10} of the number of ratings) and average rating, with product fixed effects, from the two-country dataset. Model estimations show a negative effect of popularity on rating values, and a small positive effect of the prior score.

	Germany		France	
	Model 1	Model 2	Model 3	Model 4
Log_{10} (1+# of ratings$_{i,t-1}$)	-0.18***	-0.17***	-0.048	-0.093*
	[-0.22,-0.13]	[-0.22,-0.13]	[-0.11,0.010]	[-0.17,-0.021]
Average Rating$_{i,t-1}$		0.066*		0.10**
		[0.013,0.12]		[0.026,0.18]
Log-likelihood	-85,163	-66,656	-33,400	-30,397
Number of items	749	622	748	576
Number of observations	54,079	41,465	21,328	19,195
Age and Year Fixed Effects	Yes	Yes	Yes	Yes
Product Fixed Effects	Yes	Yes	Yes	Yes

*Note: 95CI in brackets. *: p<.05, ** : p<.01, ***: p<.001.*

REFERENCES

Moe, W. W., & Schweidel, D. A. (2012). Online product opinions: incidence, evaluation, and evolution. *Marketing Science*, *31*(3), 372–386. http://doi.org/10.1287/mksc.1110.0662