Information Systems Research
Vol. 27, No. 1, March 2016, pp. 131–144
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)

http://dx.doi.org/10.1287/isre.2015.0617

Research Note

When Do Consumers Value Positive vs. Negative Reviews? An Empirical Investigation of Confirmation Bias in Online Word of Mouth

Dezhi Yin
Trulaske College of Business, University of Missouri, Columbia, Missouri 65211, yind@missouri.edu

Sabyasachi Mitra, Han Zhang
Scheller College of Business, Georgia Institute of Technology, Atlanta, Georgia 30308
{saby.mitra@scheller.gatech.edu, han.zhang@scheller.gatech.edu}

In the online word-of-mouth literature, research has consistently shown that negative reviews have a greater impact on product sales than positive reviews. Although this negativity effect is well documented at the product level, there is less consensus on whether negative or positive reviews are perceived to be more helpful by consumers. A limited number of studies document a higher perceived helpfulness for negative reviews under certain conditions, but accumulating empirical evidence suggests the opposite. To reconcile these contradictory findings, we propose that consumers can form initial beliefs about a product on the basis of the product's summary rating statistics (such as the average and dispersion of the product's ratings) and that these initial beliefs play a vital role in their subsequent evaluation of individual reviews. Using a unique panel data set collected from Apple's App Store, we empirically demonstrate confirmation bias—that consumers have a tendency to perceive reviews that confirm (versus disconfirm) their initial beliefs as more helpful, and that this tendency is moderated by their confidence in their initial beliefs. Furthermore, we show that confirmation bias can lead to greater perceived helpfulness for positive reviews (positivity effect) when the average product rating is high, and for negative reviews (negativity effect) when the average product rating is low. Thus, the mixed findings in the literature can be a consequence of confirmation bias. This paper is among the first to incorporate the important role of consumers' initial beliefs and confidence in such beliefs (a fundamental dimension of metacognition) into their evaluation of online reviews, and our findings have significant implications for researchers, retailers, and review websites.

Keywords: positive–negative asymmetry; negativity effect; positivity effect; confirmation bias; confidence in beliefs; metacognition; online word of mouth; product review; review rating; review helpfulness
History: Anindya Ghose, Senior Editor; Michael Zhang, Associate Editor. This paper was received on January 4, 2013, and was with the authors 17 months for 3 revisions. Published online in *Articles in Advance* February 19, 2016.

Introduction

Product reviews by consumers play a vital role in electronic commerce. A distinctive feature of product reviews is the numeric rating assigned by the review to the product, typically in a star format ranging from one star (very negative) to five stars (very positive). The average and other summary statistics of these numeric ratings are often displayed prominently by review websites, in addition to the individual reviews. Not surprisingly, therefore, ratings have a substantial impact on product sales (Basuroy et al. 2003, Chevalier and Mayzlin 2006, Duan et al. 2008). Notably, empirical studies have usually found that negative ratings hurt sales to a greater extent than positive ratings help sales in diverse product categories (Basuroy et al. 2003, Cao et al. 2011, Chevalier and Mayzlin 2006). An explanation for this "negativity" effect is that from an evolutionary standpoint, humans are more alert to risks in the environment because such risks have been more critical to our survival (Vaish et al. 2008).

Whereas the impact of reviews on sales is of great interest to manufacturers and retailers, another interesting question is whether positive or negative reviews are more helpful to consumers. Review helpfulness is the extent to which an online review is perceived by consumers to facilitate their decision making (Mudambi and Schuff 2010, Yin et al. 2014). We may expect that the negativity effect described earlier will also apply here, and that consumers will find negative reviews more helpful than positive reviews because negative reviews inform consumers about the risks in product

purchase and use. However, studies examining the effect of review ratings on review helpfulness have produced mixed results. Although some studies have found that negative reviews are considered more helpful under certain circumstances (e.g., Sen and Lerman 2007, Zhang et al. 2010), accumulating empirical evidence suggests the opposite (Mudambi and Schuff 2010, Pan and Zhang 2011, Scholz and Dorner 2013).

To reconcile these contradictory findings, we propose that an investigation into the way individual reviews are evaluated by the consumer needs to take into account consumers' initial beliefs about the product formed on the basis of summary rating statistics. This notion is in line with a stream of research in consumer decision making that suggests that consumers are influenced by their prior beliefs and expectations when evaluating new information (Alba et al. 1994). In our context, most online review websites display the summary statistics of products' ratings prominently, such as the average and distribution of their ratings. In particular, consumers typically see a product's average rating *before* they dig deeper and read specific reviews of the product, so the average rating provides a basis for consumers to form an initial belief about the product. This initial belief may subsequently influence how consumers read and evaluate individual reviews of the product in systematic ways (Cheung et al. 2009).

Specifically, we propose that consumers, as they make sense of and cognitively process individual reviews, exhibit confirmation bias—a tendency to prefer information that confirms their initial beliefs (Klayman and Ha 1987, Trope and Bassok 1982). Because consumers are likely to form their initial belief about a product based on its average rating, confirmation bias would predict that individual reviews that deviate more from this baseline (i.e., the product's average rating) will be perceived as less helpful. Extending the confirmation bias literature, we also propose that confirmation bias will be attenuated when people are less certain about their initial beliefs (Risen and Gilovich 2007), such as for products with a high dispersion of ratings (Rucker et al. 2014). Furthermore, we argue that an important consequence of confirmation bias is that the direct effect of review rating on review helpfulness that has been studied extensively in the literature will depend on the product's average rating (the basis of their initial belief). When the average product rating is high (e.g., four stars), positive reviews generally deviate less from the average (by zero or one star) than negative reviews (by two or three stars) because ratings are constrained to be between one and five stars, and confirmation bias predicts higher perceived helpfulness for positive reviews compared to negative reviews (positivity effect). Similarly, we argue for a negativity effect when the average product rating is low (e.g., two stars) and a lack of positive–negative asymmetry when the average product rating is at the midpoint (three stars). Thus, the positive–negative asymmetry can be a consequence of confirmation bias, and the effect of consumers' initial beliefs can lead to the contradictory findings in the literature.

To test these ideas, we collected a unique panel data set from Apple's App Store for a period of 62 days. Apps play an important role in how consumers interact with various entities on the Internet (Ghose and Han 2014). Our data set tracks "helpful" and "not helpful" votes cast by readers on a daily basis for 106,045 reviews of 505 popular apps. We augment a large-scale cross-sectional analysis of reviews that did not receive additional votes in our study period with an analysis of a smaller number of votes cast during the study period that can better account for potential endogeneity through panel data methods. Finally, since not all reviews are voted on by consumers, we correct for potential selection bias by matching voted with similar nonvoted reviews that appeared on the same webpage on the same day, and by incorporating latent (unobserved) variables in the model that affect both the selection and the likelihood of a positive vote. We find consistent support for our predictions based on confirmation bias and its consequences.

We make three key contributions to the literature. First, we examine the critical role of initial beliefs about the product, and we empirically document a confirmation bias in the cognitive processing of reviews by consumers. Note that it is also possible to theoretically argue a disconfirmation bias since consumers may find reviews that deviate from their initial beliefs more surprising, attention grabbing, and, thus, more informative (Helson 1964, Sherif and Sherif 1967). We believe that this makes our empirical findings more interesting and less a priori obvious. Second, we explore how confidence in initial beliefs, one of the fundamental dimensions of metacognition (i.e., thinking about ones' own thoughts and beliefs; see Petty et al. 2007), affects the evaluation of reviews by consumers, and we show that confirmation bias is attenuated when confidence in the initial belief is weak. Third, our findings can potentially reconcile the mixed evidence about positive–negative asymmetry in the literature. Consistent with confirmation bias, we empirically demonstrate a positivity effect when the average product rating is high and a negativity effect when the average product rating is low. Thus, the positive–negative asymmetry often studied in the literature can be a consequence of confirmation bias (for similar arguments, see Pan and Zhang (2011).

Research Framework and Hypotheses

Confirmation Bias

An emerging stream of research in information systems examines the perceived helpfulness of online

reviews and its antecedents (Ghose and Ipeirotis 2011, Mudambi and Schuff 2010, Pan and Zhang 2011). Confirmation bias—a tendency of humans to overweigh information that confirms (versus disconfirms) their initial beliefs and positions (Nickerson 1998)—can have a significant effect on the perceived helpfulness of individual reviews. There is evidence in many contexts that humans tend to prefer information that confirms their initial beliefs, hypotheses, and conjectures (Klayman and Ha 1987, Trope and Bassok 1982). According to cognitive dissonance theory, humans experience psychological discomfort when faced with evidence that contradicts their prior beliefs (Festinger 1962, Swann et al. 1987), and they depreciate disconfirmatory evidence to reduce such discomfort and maintain consistency (Darley and Gross 1983).

Review websites typically display the ratings of products at two different levels—the individual review level and the aggregated product level (Qiu et al. 2012). At the product level, aggregated information cues, such as the average and distribution of product ratings consolidated from individual review ratings, are prominently displayed in almost all online review platforms. The average rating of the product reflects aggregated evaluation of the product's quality by consumers who have already purchased the product. It serves as an easily accessible and salient signal for prospective consumers to form an initial belief about the product before they browse individual reviews (Sun 2012). A confirmation bias can occur in the evaluation of individual online reviews (Cheung et al. 2009), because information in reviews that confirms consumers' initial beliefs about the product can cause less psychological discomfort than information that contradicts their initial beliefs.

Hypothesis 1 (H1). *The deviation of a review rating from the product's average rating* (*i.e.*, *rating deviation*) *has a negative effect on the perceived helpfulness of the review.*

Confidence in Initial Beliefs

Additionally, the extent of confirmation bias can depend on the confidence of consumers in their initial beliefs. Confidence in beliefs refers to the extent of perceived certainty that their beliefs are accurate (Smith and Swinyard 1988). Recent experimental evidence in other contexts indicates that confirmation bias may be less pronounced when attitudes and beliefs are weaker and more uncertain (Fischer et al. 2010, Park et al. 2013). As people become less confident and more uncertain about their initial beliefs, they will experience a lower level of psychological discomfort when they encounter disconfirmatory information, thus decreasing the extent of confirmation bias (Hart et al. 2009). As an aside, confidence is one of the fundamental dimensions of metacognition, namely, secondary cognition: following first-order thoughts involving people's initial association of some object (e.g., a product) with some attribute (e.g., its quality), people can generate second-order thoughts that reflect on the first-level thoughts (e.g., "Is this evaluation accurate?") (Petty et al. 2007). Metacognition can magnify or attenuate first-order thoughts, but little is known about its role in online word of mouth.

In the online reviews context, the dispersion (measured through the standard deviation) of ratings reflects the consensus among reviewers and provides review readers with information on how "accurate" the average ratings are. A high dispersion of ratings indicates low agreement among reviewers, whereby the opinions of different reviewers about the product are diverging (Moe and Trusov 2011). Although high dispersion of ratings can be caused by several factors, such as diversity in consumers' tastes or product differentiation (Clemons et al. 2006, He and Bond 2015, Sun 2012), lower consensus leads consumers to be less confident in the validity of the average ratings and less certain of their initial beliefs (see Petrocelli et al. 2007). Thus, we expect confirmation bias to be attenuated when the dispersion of ratings for the product is higher.[1]

Hypothesis 2 (H2). *Confirmation bias* (*the negative effect of rating deviation on review helpfulness*) *will be weaker for products that have a higher dispersion of ratings.*

Positive–Negative Asymmetry

The relationship between review rating and review helpfulness—whether positive or negative reviews are perceived to be more helpful by consumers—is a phenomenon termed as positive–negative asymmetry in the broader literature (Baumeister et al. 2001). Empirical evidence concerning the impact of review rating on review helpfulness is inconclusive, with some studies reporting that negative reviews are rated as more helpful than positive reviews (e.g., Sen and Lerman 2007, Zhang et al. 2010), and others reporting that positive reviews are rated as more helpful than negative reviews (e.g., Korfiatis et al. 2012, Mudambi and Schuff 2010, Pan and Zhang 2011, Scholz and Dorner 2013). A probable reason for these contradictory findings is the impact of product-level summary statistics of ratings, which may shape the consumers' initial beliefs about the product even before they read and evaluate individual reviews.

Specifically, confirmation bias has important implications for positive–negative asymmetry, which has

[1] As supporting evidence, we also show that confirmation bias is weaker when the average rating of the product is not displayed by the review platform (e.g., when there are too few reviews for the product) and confidence in the initial belief is weaker as a result. For details, see Online Appendix B (available as supplemental material at http://dx.doi.org/10.1287/isre.2015.0617).

received intense interest in the online word-of-mouth literature. Consider a product that has a high average rating (e.g., four stars). Because the ratings assigned by a review are constrained between one and five stars, positive reviews with four or five stars deviate less from the average rating than negative reviews with one or two stars (i.e., deviating from the average by zero or one star versus deviating by two or three stars). Consequently, confirmation bias predicts that positive reviews will be perceived to be more helpful than negative reviews for such products, and the effect of review rating on review helpfulness will be positive. On the other hand, consider products with low average ratings (e.g., two stars). In such cases, negative reviews with one or two stars deviate less from the average rating than positive reviews with four or five stars (i.e., deviating from the average by zero or one star versus deviating by two or three stars), and confirmation bias predicts that negative reviews will be perceived to be more helpful than positive reviews, and the effect of the review rating on review helpfulness will be negative. A similar reasoning will also predict that for products with an average rating of three stars (midpoint of the range), there should be no relationship between review rating and perceived helpfulness of the review because positive and negative reviews deviate similarly from the average rating.

Hypothesis 3 (H3). *The effect of review rating on review helpfulness will be* (i) *positive for products with a high average rating*, (ii) *negative for products with a low average rating*, *and* (iii) *not significantly different from zero for products whose average rating is in the middle of the scale*.

Note that it is also possible to argue a disconfirmation bias because consumers may find reviews that deviate from the average product ratings more surprising and informative (Helson 1964, Sherif and Sherif 1967), and consequently more helpful. A disconfirmation bias would also lead to predictions that are the opposite of those in Hypothesis H3. We believe that this empirical tension makes our results more interesting, insightful, and less a priori obvious.

Data and Analysis

Data Collection and Empirical Strategy

We collected daily data on reviews from Apple's App Store from July 1 through August 31, 2013 (our study period of 62 days). We began by identifying and tracking 538 apps that appeared in the top-100 overall rankings in Apple's App Store at least once during June 2013. Of these, 505 apps had at least one review by the end of our study period. After discarding non-English characters from reviews and eliminating a few reviews that had no text, these 505 apps had 106,045 reviews that received at least one vote by the end of the period. As explained in the next paragraph, our primary data set consists of downloaded data on these 106,045 reviews and 505 apps on a daily basis during the study period of 62 days. Of these 106,045 reviews, 97,623 reviews had at least one vote at the beginning of our study period, whereas the remaining reviews were added or voted on for the first time during our study period. However, the likelihood of receiving a vote decreases quickly with the age of a review. Thus, during our study period, only 3,291 reviews for 267 apps received a total of 8,006 votes.

For each of the 106,045 reviews in our data set, we collected the following information that did not change over time: the numerical rating assigned by the review to the product, the review text, and the date the review was originally posted. In addition, the following information about the review changed over time and was tracked daily: the number of "helpful" votes cast every day and the number of "not helpful" votes cast every day for the review by users. We also collected data on each of the 505 apps in our sample. The following data on each app remained invariant over time: the app category, the date the app was initially launched, whether the app has an iPad version, the number of apps by the developer at the beginning of our study period, and the file size of the app. In addition, the following app-level information changed over time and was tracked daily: the overall ranking of the app, the average rating of the app, the distribution of the ratings (e.g., number of one-star ratings, two-star ratings, etc.), a count of all ratings for the app, the price of the app, and whether the app released a version update on a specific date.

We perform three types of analysis. First, we consider 95,926 reviews that existed in our data set but did not receive any new votes during the 62 days. These reviews are in *steady state* since no new votes were cast for these reviews during the 62 days. We perform cross-sectional analysis of these reviews to evaluate how the total number of "helpful" votes received by a review (as a fraction of the total votes received) is affected by the review rating and the deviation of the rating from the average rating of the app, after controlling for a large number of app and review characteristics. Second, we consider the remaining 3,291 reviews in our data set that received at least one vote during the 62 days in our study period, and we evaluate the 8,006 votes cast for such reviews through panel data methods. The unit of analysis is a vote cast by a user, and we evaluate how the likelihood of a "helpful" vote is affected by the review rating and the deviation of the rating from the average rating of the app. In this analysis, all variables are calculated at the time the vote was cast. Since more than one vote can be cast for the same review by users, we can better account for

Table 1 Variable Definitions

Variable name	Operationalization
$RVotes_{ijt}$	Cumulative number of votes for review j of app i at time t
$RPVotes_{ijt}$	Cumulative number of positive (helpful) votes for review j of app i at time t
V_{ijk}	Equals 1 if vote k for review j of app i is positive (or "helpful"), 0 otherwise
$AMRating_{it}$	Average rating of app i at time t
$RRating_{ij}$	Star rating assigned by review j to app i
$RDev_{ijt}$	Absolute difference between review rating ($RRating_{ij}$) and the average rating of the app ($AMRating_{it}$)
$ADisp_{it}$	Standard deviation of the ratings corresponding to app i at time t
$ARank_{it}$	Rank of app i at time t based on the number of downloads
$APrice_{it}$	Price per download of app i at time t
$AUpd_{it}$	Equals 1 if app i released an update (new version) at time t, 0 otherwise
$ARCount_{it}$	Cumulative number of ratings for app i at time t
$AiPad_i$	Equals 1 if app i is also available in an iPad version, 0 otherwise
$ADevN_i$	Number of other apps developed by the maker of app i
$ASize_i$	Size of app i (in MB)
$RLength_{ij}$	Number of words in review j of app i
$RDiff_{ij}$	Gunning–Fog index of the reading difficulty of review j of app i
$REmo_{ij}$	Percentage of words in review j of app i that indicate either positive or negative emotions
$RCog_{ij}$	Percentage of words in review j of app i that are related to reasoning or cognitive mechanisms
$RDays_{ijt}$	Age of review j of app i at time t

correlated error terms and for unobserved factors of a review that affect the likelihood of a positive vote through panel data methods.

Finally, since only a small subset of reviews receive votes from the users, the vote-level analysis can suffer from selection bias because unobserved factors that affect the selection of a review can also affect the likelihood of a positive vote. To correct for this bias, we identify all reviews that appeared on the same Web page on the same day as a voted review in the default sort order in the App Store but did not receive a vote from users. When a user voted on a review, it is very likely that she saw all reviews on the same page, and this smaller subset of reviews for the same app is an ideal comparison set for the voted reviews. We use an extension of the generalized linear model that allows latent (unobserved) variables and mixed effects described in Rabe-Hesketh et al. (2002a) to correct for the selection bias in our estimates. We obtained consistent results through the cross-sectional and vote-level analysis, and after accounting for selection bias.

Variable Definitions

Table 1 describes the variables in the empirical analysis. Table 2 presents summary statistics and shows the correlations between selected variables in our data set. In the variable descriptions that follow, i indexes an app, j indexes a review for an app, k indexes a vote for review, and t indexes the time (day) in our study period. We use $RVotes_{ijt}$ to denote the total (cumulative) number of votes cast by readers for review j of app i at time t, and use $RPVotes_{ijt}$ to denote the total (cumulative) number of positive ("helpful") votes cast by users for review j of app i at time t. The dependent variable in our cross-sectional analysis of 95,926 reviews in steady state is $RPVotes_{ijt}$ at the end of the study period ($t =$ August 31, 2013). In the vote-level analysis, V_{ijk} is an indicator variable that is 1 if vote k for review j of app i is a positive (or "helpful") vote and 0 if the vote is a negative (or "not helpful") vote. The dependent variable for the vote-level analysis is V_{ijk}.

The variable $AMRating_{it}$ is the average rating (between one and five stars) of app i at time t. As more reviews are posted, the $AMRating_{it}$ variable changes over time for an app, although such changes are greater for new apps or when apps undergo version updates. The variable $RRating_{ij}$ is the rating (an integer between one and five stars) assigned by review j to app i, and it remains invariant over time. The rating deviation ($RDev_{ijt}$) is the absolute value of the deviation of the rating assigned by a review from the average rating of the app (i.e., $RDev_{ijt} = abs(RRating_{ij} - AMRating_{it})$). The dispersion of review ratings for an app ($ADisp_{it}$) is the standard deviation of the ratings corresponding to app i at time t. Our primary independent variables and moderators are $RDev_{ijt}$, $RRating_{ij}$, $ADisp_{it}$, and $AMRating_{it}$.

We use several control variables in our analysis. The following variables are defined for each app, and they can indirectly affect the likelihood of a "helpful" vote for a review through their effect on user perceptions about the app. As a proxy for daily app sales, $ARank_{it}$ is the overall rank of app i at time t based on the number of downloads. The variable $APrice_{it}$ is the price per download of app i at time t. The variable $AUpd_{it}$ is set to 1 if app i released an update (new version) at time t and 0 otherwise. The variable $ARCount_{it}$ is the cumulative number of ratings for app i at time t, and it is a measure of the overall consumer interest expressed in app i. The variable $AiPad_i$ is set to 1 if the app is also

Table 2 Descriptive Statistics and Correlations

At review level for reviews in steady state

Variable	*N*	Mean	Std. dev.	Min.	Max.	1	2	3	4	5	6	7	8	9
1 $RVotes_{ijT}$	95,926	3.67	13.10	0	1,557	1								
2 $RPVotes_{ijT}$	95,926	4.61	15.98	1	1,628	0.96	1							
3 $RRating_{ij}$	95,926	3.32	1.76	1	5	−0.03	−0.05	1						
4 $RDev_{ijT}$	94,263	1.48	1.14	0	4	0.03	0.05	−0.78	1					
5 $RLength_{ij}$	95,926	33.13	41.72	1	1,105	0.09	0.08	−0.13	0.08	1				
6 $RDiff_{ij}$	95,926	6.72	4.19	0.4	155.6	0.04	0.03	−0.06	0.03	0.30	1			
7 $REmo_{ij}$	95,926	11.92	15.60	0	100	−0.04	−0.04	0.29	−0.20	−0.25	−0.27	1		
8 $RCog_{ij}$	95,926	14.91	9.95	0	100	0.01	0.01	−0.07	0.02	0.11	0.18	−0.23	1	
9 $RDays_{ijT}$	95,924	614.53	441.86	62	1,878	0.08	0.10	0.02	0.02	0.12	0.04	−0.09	0.03	1

At review vote level

Variable	*N*	Mean	Std. dev.	Min.	Max.	1	2	3	4	5	6	7	8
1 V_{ijk}	8,006	0.76	0.43	0	1	1							
2 $RRating_{ij}$	8,006	2.79	1.79	1	5	0.08	1						
3 $RDev_{ijt}$	7,889	1.79	1.17	0	4	−0.10	−0.75	1					
4 $RLength_{ij}$	8,006	45.75	62.68	1	863	0.06	−0.13	0.11	1				
5 $RDiff_{ij}$	8,006	7.39	4.10	0.4	43.9	0.04	−0.08	0.06	0.27	1			
6 $REmo_{ij}$	8,006	9.77	11.67	0	100	0.01	0.31	−0.19	−0.19	−0.25	1		
7 $RCog_{ij}$	8,006	15.51	9.06	0	100	0.02	−0.01	0.00	0.06	0.15	−0.22	1	
8 $RDays_{ijt}$	8,006	69.45	142.29	0	1,747	0.02	0.06	−0.05	0.02	0.05	−0.06	0.07	1

available in an iPad version, 0 otherwise. The variable $ADevN_i$ is the number of other apps developed by the maker of app i at the start of our study period, and is a measure of the experience level of the developer and the size of their operations. The variable $ASize_i$ is the size (in MB) of app i and it can affect the number of downloads of app i.

The following control variables are defined for each review of an app and they can directly affect the perceived helpfulness of the review. The variable $RLength_{ij}$ is the number of words in review j of app i and is a measure of the amount of detail provided by the review. The variable $RDiff_{ij}$ is the Gunning–Fog index that measures the reading difficulty of review j of app i, and it indicates the number of years of formal education needed to understand the text of the review on a first reading (Gunning 1968). The variable $REmo_{ij}$ is the percentage of words in the review that indicate either positive or negative emotions, and $RCog_{ij}$ is the percentage of words in the review that are related to reasoning or cognitive mechanisms. Both $REmo_{ij}$ and $RCog_{ij}$ are calculated using the text analysis software Linguistic Inquiry and Word Count (Pennebaker et al. 2007). The variable $RDays_{ijt}$ is the age of review j of app i at time t, and this variable changes over time.

Cross-Sectional Analysis of Reviews

In this analysis, we focus on the 95,926 *steady state* reviews in our data set that did not receive any additional votes during the 62 days in our study period, and we construct a data set that contains, for each such review, the number of "helpful" ($RPVotes_{ijt}$) and total votes ($RVotes_{ijt}$) received by the review since its inception, as well as the independent and control variables described earlier. All variables are calculated for the last day in our study period and for the current version of the app. In Equations (1) and (2), ϕ is the logit function, U_c is the fixed effects intercept for the app category, and α_i is the random intercept for app i. Following prior research (Forman et al. 2008), we log transformed count variables, including app ranking, number of ratings, number of apps from the developer, app size, review length, and review age. The unit of analysis is a single review, and T represents the last day in our study period ($T =$ August 31, 2013). Each vote is a Bernoulli trial with two outcomes (similar to a coin toss) with probability parameter θ_{ij}, which is invariant over time in the cross-sectional analysis. Thus, the total number of positive votes ($RPVotes_{ijT}$) for a review is binomially distributed with probability parameter θ_{ij} and $RVotes_{ijT}$ trials

$$\begin{aligned}\varnothing(\theta_{ij}) = {} & \beta_0 + \beta_1 ARank_{iT} + \beta_2 APrice_{iT} + \beta_3 AUpd_{iT} \\ & + \beta_4 ARCount_{iT} + \beta_5 AiPad_i + \beta_6 ADevN_{iT} \\ & + \beta_7 ASize_i + \beta_8 ADisp_{iT} + \beta_9 AMRating_{iT} \\ & + \beta_{10} RLength_{ij} + \beta_{11} RDiff_{ij} + \beta_{12} REmo_{ij} \\ & + \beta_{13} RCog_{ij} + \beta_{14} RDays_{ijT} + \beta_{15} RDev_{ijT} \\ & + \beta_{16} RDev_{ijT} \times ADisp_{iT} + U_c + \alpha_i + \epsilon_{ij}, \quad (1)\end{aligned}$$

$$RPVotes_{ijT} \sim \text{Binomial}[RVotes_{ijT}, \theta_{ij}]. \quad (2)$$

We estimate (1) and (2) using maximum likelihood estimation and *mixed effects generalized linear models*

Table 3 Cross-Sectional Analysis of Reviews

DV: $\varnothing(\theta_{ij})$ in (1)	Model 1	Model 2	Model 3	Model 4	Model 5
$RLength_{ij}$ (Ln)	0.146***	0.159***	0.164***	0.167***	0.171***
	(0.017)	(0.019)	(0.018)	(0.018)	(0.018)
$RDiff_{ij}$	0.007**	0.004	0.005	0.005	0.005
	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)
$REmo_{ij}$	0.005***	0.001	0.001	0.001	0.001
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
$RCog_{ij}$	0.001	0.001	0.001	0.001	0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$RDays_{ijT}$ (Ln)	−0.730***	−0.754***	−0.734***	−0.777***	−0.744***
	(0.069)	(0.070)	(0.068)	(0.070)	(0.069)
$ARank_{iT}$ (Ln)	0.122***	0.128***	0.130***	0.123**	0.130***
	(0.047)	(0.048)	(0.047)	(0.048)	(0.048)
$APrice_{iT}$	−0.029	−0.032*	−0.032*	−0.032*	−0.033*
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
$AUpd_{iT}$	−0.999***	−0.997***	−1.077***	−1.039***	−1.003***
	(0.273)	(0.211)	(0.252)	(0.198)	(0.214)
$ARCount_{iT}$ (Ln)	0.021	0.044	0.046	−0.021	−0.031
	(0.036)	(0.036)	(0.036)	(0.118)	(0.118)
$AiPad_{i}$	0.008	−0.027	−0.021	0.079**	0.073**
	(0.117)	(0.118)	(0.117)	(0.038)	(0.037)
$ADevN_{i}$ (Ln)	0.074**	0.075**	0.072*	0.039	0.047
	(0.037)	(0.037)	(0.037)	(0.036)	(0.036)
$ASize_{i}$ (Ln)	−0.106*	−0.093	−0.094*	−0.101*	−0.088
	(0.058)	(0.058)	(0.056)	(0.059)	(0.058)
$AMRating_{iT}$	0.340**	0.382**	0.355**	0.288*	−0.126
	(0.155)	(0.176)	(0.179)	(0.162)	(0.165)
$ADisp_{iT}$	0.220	0.367	−0.401	0.279	0.457
	(0.303)	(0.343)	(0.319)	(0.329)	(0.363)
$RDev_{ijT}$		−0.268***	−0.726***		
		(0.053)	(0.138)		
$RDev_{ijT} \times ADisp_{iT}$			0.419***		
			(0.098)		
$RRating_{ij}$				0.155***	−0.544***
				(−0.037)	(0.146)
$RRating_{ij} \times AMRating_{iT}$					0.173***
					(0.040)
Constant	3.542***	3.582***	4.328***	3.347***	4.417***
	(1.097)	(1.220)	(1.182)	(1.218)	(1.204)
Observations	86,854	86,854	86,854	86,854	86,854
Log likelihood	−78,771.6	−76,148.9	−75,440.0	−76,904.9	−75,830.7

Notes. Standard errors are in parentheses. Fixed effects at the category level and random effects at the app level are included. Ln indicates the variable is log transformed.

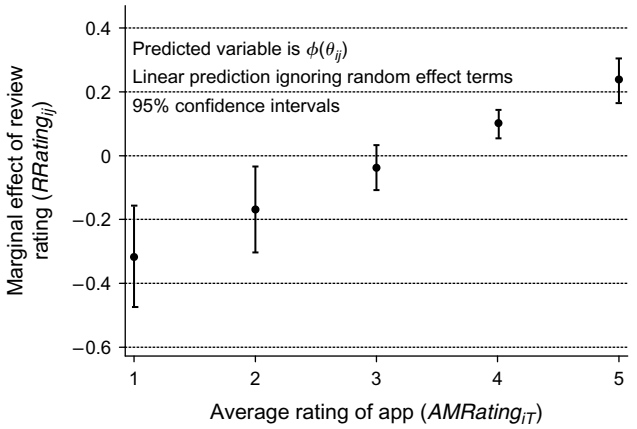
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

(Wooldridge 2010) in Stata.[2] The results are shown in Table 3. The results are based on 86,854 reviews for which all control and focal variables were available. Model 1 (Table 3) introduces the control variables. As expected, longer reviews ($RLength_{ij}$), newer reviews ($RDays_{ijT}$), and reviews for apps with higher average ratings ($AMRating_{ijT}$) are more likely to receive "helpful" votes. Model 2 (Table 3) introduces the $RDev_{ijT}$ variable. The coefficient for the $RDev_{ijT}$ variable is negative and significant ($\beta = -0.27$, $p < 0.01$), indicating that a one-star deviation of review rating from the average rating of the app decreases the odds[3] of a "helpful" vote by 24% ($e^{-0.27} = 0.76$). Thus, Hypothesis H1 is supported in the cross-sectional analysis. Model 3 (Table 3) introduces the interaction term ($RDev_{ijT} \times ADisp_{iT}$). The coefficient for the interaction term is significant and

[2] We use the meglm command in Stata with a logit link and binomially distributed dependent variable. Details of the Stata procedure appears in Rabe-Hesketh et al. (2002b).

[3] The odds of an event is the ratio $p/(1-p)$, where p is the probability of the event. With the logit link function, β is the change in Log(odds) and the corresponding change in odds of the event is e^{β}.

Figure 1 Marginal Effects in Cross-Sectional Analysis



positive ($\beta = 0.42, p < 0.01$), indicating that the negative effect of $RDev_{ijT}$ is weaker when the dispersion of ratings ($ADisp_{iT}$) is higher. Thus, Hypothesis H2 is supported in the cross-sectional analysis.

To evaluate Hypothesis H3, we modify (1) to exclude the $RDev_{ijT}$ variable and include the $RRating_{ij}$ variable and the interaction term ($RRating_{ij} \times AMRating_{iT}$). The coefficient for the $RRating_{ij}$ variable (Model 4 in Table 3) is significant and positive ($\beta = 0.16, p < 0.01$), indicating that a one-star increase in the rating assigned by a review increases the odds of a "helpful" vote for the review by 17% ($e^{0.16} = 1.17$). In Model 5 (Table 3), the coefficient for the $RRating_{ij}$ term is significant and negative ($\beta = -0.54, p < 0.01$), and the coefficient of the interaction term is positive and significant ($\beta = 0.17, p < 0.01$). Figure 1 plots the marginal effect of $RRating_{ij}$ on $\varnothing(\theta_{ij})$ at different values of the average app rating ($AMRating_{iT}$), along with the corresponding 95% confidence intervals. Figure 1 shows that there is a negativity effect when the average app rating is one or two stars (marginal effect of $RRating_{ij}$ is negative and significant), there is a positivity effect when the average app rating is four or five stars (marginal effect of $RRating_{ij}$ is positive and significant), and there is no effect of review rating when the average app rating is three stars. Thus, Hypothesis H3 is supported in the cross-sectional analysis.

Analysis of Votes

We focus next on the remaining 3,291 reviews in our data set that received at least one vote during the 62 days in our study period. In our data collection, we tracked the specific date when each vote was cast for a review. The unit of analysis here is a vote cast by a user, and we evaluate the 8,006 votes cast for 3,291 reviews through panel data methods. The dependent variable in this analysis is V_{ijk}, which indicates whether vote k for review j of app i is a "helpful" ($V_{ijk} = 1$) or "not helpful" ($V_{ijk} = 0$) vote. We use the same control and independent variables as before, with each variable now calculated on the day the vote was cast. All app-level variables and summary statistics are calculated for the current version of the app when the vote was cast, since the App Store displays summary statistics for the current app version by default. In this analysis, there can be multiple votes for the same review.

To evaluate Hypotheses H1 and H2, we estimate the following mixed effects logistic regression in Stata. The variable definitions appear earlier in the paper (see "Variable Definitions"). The term U_c is the fixed effects intercept for the app category, α_i is the random intercept for app i, and v_j is the random intercept for review j. In (3), $\varnothing$ is the logit function, and θ_{ijk} is the probability parameter such that $\mathrm{E}(V_{ijk}) = \theta_{ijk}$

$$\begin{aligned}\varnothing(\theta_{ijk}) = {} & \beta_0 + \beta_1 ARank_{it} + \beta_2 APrice_{it} + \beta_3 AUpd_{it} \\ & + \beta_4 ARCount_{it} + \beta_5 AiPad_i + \beta_6 ADevN_{it} \\ & + \beta_7 ASize_i + \beta_8 ADisp_{it} + \beta_9 AMRating_{it} \\ & + \beta_{10} RLength_{ij} + \beta_{11} RDiff_{ij} + \beta_{12} REmo_{ij} \\ & + \beta_{13} RCog_{ij} + \beta_{14} RDays_{ijt} + \beta_{15} RDev_{ijt} \\ & + \beta_{16} RDev_{ijt} \times ADisp_{it} + U_c + \alpha_i + v_j + \epsilon_{ijk}. \quad (3)\end{aligned}$$

Note that the random effect terms α_i and v_j take into account correlated error terms for votes of the same review and app appropriately (Wooldridge 2010), but they do not control for unobserved review-level characteristics that can be correlated with the independent variables (see the conditional logit analysis in Online Appendix A that controls for unobserved review-level characteristics but is based on fewer votes with significant changes in the $RDev_{ijt}$ variable). However, since α_i and v_j are random effect terms, we can include several app- and review-level control variables in the analysis that are invariant over time. The results of the analysis appear in Table 4. The results are based on 7,626 votes (of 8,006) that had all control and focal variables available. Model 1 (Table 4) introduces the control variables. Model 2 (Table 4) introduces the $RDev_{ijt}$ variable. The coefficient of $RDev_{ijt}$ is significant and negative ($\beta = -0.48, p < 0.01$), indicating that a one-star deviation of the review rating from the average app rating decreases the odds of a "helpful" vote by 38% ($e^{-0.48} = 0.62$). Thus, we find support for Hypothesis H1. Model 3 (Table 4) introduces the interaction term ($RDev_{ijt} \times ADisp_{it}$). The coefficient for the interaction term is significant and positive ($\beta = 0.55, p < 0.01$), indicating that the negative effect of rating deviation is weaker when the dispersion of ratings is higher. Thus, Hypothesis H2 is supported.[4]

[4] A few votes in our sample were cast when the corresponding app did not have summary-level statistics displayed (such as when the number of ratings for the app was below a threshold). Such votes were excluded from the analysis described here. In Online

Table 4 Analysis of Votes

DV: $\varnothing(\theta_{ijk})$ in (3)	Model 1	Model 2	Model 3	Model 4	Model 5
$RLength_{ij}$ (Ln)	0.149**	0.184***	0.189***	0.203***	0.207***
	(0.065)	(0.064)	(0.064)	(0.065)	(0.064)
$RDiff_{ij}$	0.017	0.009	0.010	0.009	0.010
	(0.016)	(0.015)	(0.015)	(0.015)	(0.015)
$REmo_{ij}$	0.020***	0.011**	0.010*	0.008	0.008
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$RCog_{ij}$	0.002	0.001	0.001	0.003	0.003
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
$RDays_{ijt}$ (Ln)	−0.065	−0.050	−0.054	−0.061	−0.049
	(0.048)	(0.047)	(0.047)	(0.047)	(0.047)
$ARank_{it}$ (Ln)	−0.040	−0.025	−0.026	−0.025	−0.019
	(0.059)	(0.059)	(0.058)	(0.059)	(0.059)
$APrice_{it}$	0.064	0.052	0.049	0.051	0.045
	(0.047)	(0.047)	(0.047)	(0.047)	(0.047)
$AUpd_{it}$	−0.458	−0.376	−0.432	−0.348	−0.379
	(0.313)	(0.314)	(0.314)	(0.316)	(0.315)
$ARCount_{it}$ (Ln)	−0.000	0.051	0.058	0.051	0.066
	(0.065)	(0.064)	(0.064)	(0.065)	(0.064)
$AiPad_{i}$	0.369	0.365	0.355	0.392	0.375
	(0.331)	(0.328)	(0.325)	(0.329)	(0.329)
$ADevN_{i}$ (Ln)	−0.181*	−0.178*	−0.178*	−0.163	−0.168*
	(0.100)	(0.099)	(0.098)	(0.099)	(0.099)
$ASize_{i}$ (Ln)	0.041	0.055	0.059	0.040	0.063
	(0.133)	(0.131)	(0.130)	(0.132)	(0.132)
$AMRating_{it}$	−0.701***	−0.558**	−0.607***	−0.735***	−1.448***
	(0.230)	(0.228)	(0.227)	(0.230)	(0.263)
$ADisp_{it}$	−1.668***	−1.304***	−2.318***	−1.397***	−1.173**
	(0.473)	(0.470)	(0.561)	(0.472)	(0.470)
$RDev_{ijt}$		−0.475***	−1.093***		
		(0.053)	(0.192)		
$RDev_{ijt} \times ADisp_{it}$			0.552***		
			(0.164)		
$RRating_{ij}$				0.344***	−0.837***
				(0.038)	(0.207)
$RRating_{ij} \times AMRating_{it}$					0.291***
					(0.051)
Constant	6.060***	5.434***	6.624***	4.509**	6.773***
	(1.785)	(1.769)	(1.801)	(1.781)	(1.828)
Observations	7,626	7,626	7,626	7,626	7,626
Log likelihood	−3,375.8	−3,333.1	−3,327.4	−3,330.6	−3,314.0

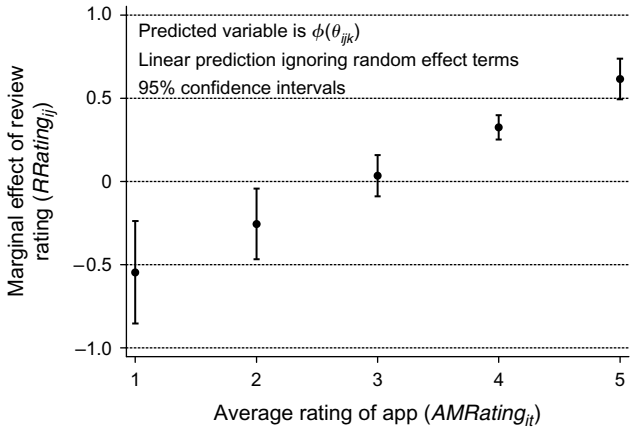
Notes. Standard errors are in parentheses. Fixed effects at the category level and random effects at the app level are included. Ln indicates the variable is log transformed.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

To evaluate Hypothesis H3, we modify (3) to exclude $RDev_{ijt}$ and include $RRating_{ij}$ and the interaction term ($RRating_{ij} \times AMRating_{it}$). In Model 4 (Table 4), the coefficient for $RRating_{ij}$ is positive and significant ($\beta = 0.34$, $p < 0.01$), indicating that a one-star increase in review rating increases the odds of receiving a helpful vote by 40% ($e^{0.34} = 1.40$). In Model 5 (Table 4), the coefficient for $RRating_{ij}$ is significant and negative ($\beta = -0.84$, $p < 0.01$), and the coefficient of the interaction term is positive and significant ($\beta = 0.29$, $p < 0.01$). Figure 2 plots the marginal effect of $RRating_{ij}$ on $\varnothing(\theta_{ijk})$ at different values of the average app rating ($AMRating_{it}$), along with the corresponding 95% confidence intervals. As in the cross-sectional analysis, there is a negativity effect when the average app rating is one or two stars, a positivity effect when the average app rating is four or five stars, and no effect when the average app rating is three stars. Thus, we find support for Hypothesis H3 in the vote-level analysis.

Appendix B, we provide an analysis of such votes to show that confirmation bias is weaker when the app-level summary statistics are not displayed and hence the consumer does not form a strong initial belief about the app. Those results also support the idea behind Hypothesis H2 that confirmation bias is weaker when the confidence in the initial belief is weak.

Figure 2 Marginal Effects in Vote-Level Analysis



Selection Bias in the Analysis of Votes

In the analysis of votes described in the previous section, only 3,291 of the 106,045 reviews in our sample received votes during the 62 days in our study period. Thus, there is a possibility of selection bias if unobserved factors that affect the selection of a review for voting also affect the likelihood of a "helpful" vote.

To correct for this bias, we construct our data set as follows. For each of the 8,006 votes in the study period, we first identify all reviews of the same app that appeared on the same Web page as a voted review on the day the vote was cast, in the default sort order of reviews in the App Store. In our data collection, we did not change the default sort order of the reviews displayed by Apple. Thus, if the reader also did not change the sort order, it is very likely that the nonvoted reviews on the same page in our data were seen by the reader when she voted on a review. We identified 31,495 such nonvoted reviews from our data. We then used the *coarsened exact matching* (CEM) procedure in Blackwell et al. (2009) and Iacus et al. (2012) to divide the reviews into groups that contained at least one voted and at least one nonvoted review for the same app (with all nonvoted reviews appearing on the same day on the same page as voted reviews within the same group), and that were closely matched on the length of the review ($RLength_{ij}$), the number of emotional words in the review ($REmo_{ij}$), the number of words indicating cognitive mechanisms ($RCog_{ij}$), and the reading difficulty of the review ($RDiff_{ij}$). We dropped groups created by the CEM procedure that did not have at least one voted and at least one nonvoted review. Thus, we have a high degree of confidence that nonvoted reviews in the selected sample were seen by readers who cast votes and were similar to voted reviews within the same group.

Our final data set contained 7,572 votes and 15,796 matched nonvoted reviews organized into 2,304 groups. Each group contains at least one voted and at least one nonvoted review, all belonging to the same app and that appeared on the same Web page in the default sort order. Let m index the records in this data set consisting of 7,572 votes and 15,796 nonvoted reviews. Let S_{ijm} be an indicator variable such that $S_{ijm}=1$ if the record m represents a vote and 0 if it represents a nonvoted review. Let g_m represent the group (described in the previous paragraph) associated with record m. We model selection bias through a latent (or unobserved) variable that affects both the selection of a review to vote and the likelihood of a positive vote if selected. The approach here is modified from Grilli and Rampichini (2007) and Rabe-Hesketh et al. (2002b). We use a probit formulation (4) for the selection stage (which reviews are selected for a vote), and a logit (5) for the primary model (whether the vote is positive or negative). We also incorporate group-level random effects in both models to capture unobserved heterogeneity across groups

$$\begin{aligned}\varphi(\pi_{ijm}) &= \alpha_0+\alpha_1 RDays_{ijt}+\alpha_2 ARank_i+\alpha_3 APrice_{it}\\ &\quad+\alpha_4 ARCount_{it}+\alpha_5 RDev_{ijt}+\alpha_6 REmo_{ij}\\ &\quad+W_{ijm}+\gamma_{g_m}+\epsilon_{ijm};\end{aligned} \tag{4}$$

$$\begin{aligned}\varnothing(\theta_{ijm}) &= \beta_0+\beta_1 AUpd_{it}+\beta_2 AiPad_i+\beta_3 ADevN_{it}\\ &\quad+\beta_4 ASize_i+\beta_5 ADisp_{it}+\beta_6 AMRating_{it}\\ &\quad+\beta_7 RLength_{ij}+\beta_8 RDiff_{ij}+\beta_9 REmo_{ij}\\ &\quad+\beta_{10} RCog_{ij}+\beta_{11} RDev_{ijt}\\ &\quad+\beta_{12} RDev_{ijt}\times ADisp_{it}\\ &\quad+U_c+\omega W_{ijm}+v_{g_m}+\delta_{ijm}.\end{aligned} \tag{5}$$

In (4), φ is the inverse of the cumulative distribution function of the standard normal distribution, and π_{ijm} is the probability parameter such that $\mathrm{E}(S_{ijm})=\pi_{ijm}$ (standard probit model). In (5), $\varnothing$ is the logit function, and θ_{ijm} is the probability parameter such that $\mathrm{E}(V_{ijm})=\theta_{ijm}$ (logit model), defined only when $S_{ijm}=1$. For a vote, all variables are calculated at the time the vote was cast. Recall that nonvoted reviews in our sample appeared on the same page as a voted review on the day the vote was cast. Thus, for nonvoted reviews in the selection model, all variables are calculated on the date of the corresponding vote that appeared on the same page in the default sort order. The term W_{ijm} represents a latent (unobserved) variable that affects both the selection of the review for a vote and the likelihood of receiving a "helpful" vote when selected. If the unobserved components in the two models are not correlated, the estimated parameter ω for this variable in (5) should not be significantly different from zero. The terms γ_{g_m} and v_{g_m} are random effects in (4) and (5) based on the groups defined earlier.

The rationale behind the different variables included in (4) and (5) is as follows. We assume that the likelihood of selecting a review to cast a vote on (among

many that appear on the same page) is affected by factors related to the popularity of the app (its rank, its price, and the number of reviews) as well as review-level factors such as the age of the review, the deviation of the review rating from the average rating of the app, and the percentage of emotional words in the review. Once a review has been selected for a vote, we assume that the likelihood of a "helpful" vote is not affected by the popularity of the app or the age of the review. Without loss of generality, the variance of W_{ijm} is set to 1 for identification.

We estimate (4) and (5) jointly through maximum likelihood estimation using the generalized linear latent and mixed models (GLLAMM) procedure in Stata. The advantage of the GLLAMM procedure is that it can incorporate fixed effects (U_c based on app category), random effects (γ_{g_m} and v_{g_m} based on the groups), latent variables (W_{ijm} in both equations), and separate link functions and distributions for the dependent variables (probit for the selection model and logit for the main model). However, since no closed form solutions exist for the likelihood function, GLLAMM relies on numerical integration and is therefore extremely time consuming and sometimes fails to converge in reasonable time. Other alternative estimation methods are possible, such as Bayesian Markov Chain Monte Carlo and maximum simulated likelihood (Grilli and Rampichini 2007).

The results of the analysis are presented in Table 5. Panel A of Table 5 shows the estimates from the selection model. We find that older reviews are less likely to receive a vote, and deviation of the review rating from the average rating of the app marginally increases the likelihood of a vote. Since we have carefully matched the voted and nonvoted reviews on observed characteristics, it appears that the selection of reviews for voting is otherwise random for similar reviews on the same page.[5] Panel B of Table 5 shows the estimates of the main model. Model 1 in Table 5 introduces the control variables, whereas Model 2 introduces the $RDev_{ijt}$ variable. The coefficient for the $RDev_{ijt}$ variable is significant and negative ($\beta = -0.33, p < 0.01$), indicating support for Hypothesis H1. In Model 3 of Table 5, the coefficient for the $RDev_{ijt}$ variable is significant and negative ($\beta = -1.20, p < 0.01$), and the coefficient for the interaction term ($RDev_{ijt} \times ADisp_{it}$) is significant and positive ($\beta = 0.77, p < 0.01$), providing support for Hypothesis H2. In Model 5 in Table 5, the coefficient for the $RRating_{ij}$ variable is negative but not significant ($\beta = -0.07, p > 0.1$), but the coefficient for the interaction term is significant and positive ($\beta = 0.09, p < 0.01$), consistent with Hypothesis H3. The coefficient of the unobserved latent variable (W_{ijm}) in (5) is not significant in any of the models. Unlike earlier, it is not possible to compute the confidence intervals of the marginal effects of the variables through the GLLAMM procedure.

[5] A simple logit analysis *without* the group-level effects is shown in Online Appendix C. It shows that newer reviews and reviews with fewer emotional words are more likely to receive votes from consumers. Also, reviews whose ratings deviate more from the average app rating are more likely to receive votes (but less likely to receive positive votes, as shown in Table 5, Panel B).

Discussions and Implications

Building on the confirmation bias literature, we demonstrate that individual reviews whose ratings deviate from product average ratings—the basis for consumers to form initial beliefs—are perceived as less helpful by consumers, and that this confirmation bias is attenuated when the confidence in the initial belief is weak (such as when the dispersion of ratings for the app is high or when summary statistics are not available for the app). This paper is among the first attempts at incorporating the role of initial beliefs (see also Cheung et al. 2009, Qiu et al. 2012) and confidence in such beliefs into consumer perceptions of online word of mouth. We also demonstrate a higher perceived helpfulness for positive reviews compared with negative reviews (positivity effect) when the average product rating is high, an opposite negativity effect when the average product rating is low, and a lack of positive–negative asymmetry when the average product rating is at the midpoint. Thus, the positive–negative asymmetry that has been extensively studied in the literature can be a consequence of confirmation bias (for similar arguments, see Pan and Zhang 2011), and the effect of consumers' initial beliefs can lead to the contradictory findings in the literature. For example, in controlled experiments where participants' initial beliefs are absent as they evaluate individual reviews (e.g., Sen and Lerman 2007, Zhang et al. 2010), a negativity effect is likely, given our evolutionary conditioning to be more alert to risks in the environment (Vaish et al. 2008). By contrast, in empirical studies utilizing real-world data sets (e.g., Korfiatis et al. 2012, Mudambi and Schuff 2010, Pan and Zhang 2011, Scholz and Dorner 2013), a positivity effect is likely because the average rating is high for most products (Chevalier and Mayzlin 2006).

Our findings can help to improve review websites that want to better inform consumers in their decision making. To reduce review readers' confirmatory tendencies and to focus their attention on content quality, it may be advisable to tweak the ways that helpfulness votes are solicited. For example, a more adequate question in this case may be, does this review provide helpful content? or is this review accurate and informative? rather than simply asking, is the review helpful? Another approach is to promote certain negative reviews even if they are not voted as helpful as positive reviews. For instance, Amazon lists the

Table 5 Selection Bias in the Analysis of Votes

	Model 1	Model 2	Model 3	Model 4	Model 5
			Panel A: Selection model		
$RDays_{ijt}$ (Ln)	−1.078*	−1.127	−1.032**	−1.184	−1.973
	(0.598)	(0.721)	(0.519)	(0.877)	(5.095)
$ARank_{it}$ (Ln)	0.063	0.065	0.060	0.069	0.115
	(0.041)	(0.048)	(0.037)	(0.056)	(0.299)
$APrice_{it}$	−0.021	−0.022	−0.020	−0.023	−0.039
	(0.017)	(0.019)	(0.015)	(0.022)	(0.103)
$ARCount_{it}$ (Ln)	0.032	0.033	0.031	0.035	0.059
	(0.027)	(0.030)	(0.024)	(0.034)	(0.155)
$RDev_{ijt}$	0.140*	0.146	0.134*	0.154	0.256
	(0.084)	(0.099)	(0.074)	(0.119)	(0.663)
$REmo_{ij}$	−0.002	−0.002	−0.002	−0.003	−0.004
	(0.004)	(0.004)	(0.003)	(0.004)	(0.013)
Constant	1.475	1.541	1.412*	1.621	2.699
	(0.843)	(1.008)	(0.735)	(1.221)	(6.981)
			Panel B: Primary model		
$RLength_{ij}$ (Ln)	0.146***	0.177***	0.180***	0.218***	0.215***
	(0.049)	(0.050)	(0.050)	(0.051)	(0.051)
$RDiff_{ij}$	0.020*	0.015	0.017	0.013	0.013
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
$REmo_{ij}$	0.018***	0.013***	0.011***	0.007*	0.008*
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
$RCog_{ij}$	0.006	0.006	0.006	0.006	0.007
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
$AUpd_{it}$	−0.180	−0.201	−0.251	−0.210	−0.227
	(0.214)	(0.217)	(0.220)	(0.219)	(0.220)
$AiPad_{i}$	0.493***	0.540***	0.548***	0.548***	0.560***
	(0.109)	(0.109)	(0.110)	(0.109)	(0.110)
$ADevN_{i}$ (Ln)	−0.156***	−0.161***	−0.167***	−0.135***	−0.143***
	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)
$ASize_{i}$ (Ln)	0.119***	0.174***	0.172***	0.159***	0.174***
	(0.041)	(0.042)	(0.043)	(0.042)	(0.043)
$AMRating_{it}$	−0.508***	−0.528***	−0.546***	−0.524***	−0.800***
	(0.105)	(0.105)	(0.105)	(0.107)	(0.149)
$ADisp_{it}$	−0.083	−0.057	−1.421***	0.156	0.118
	(0.216)	(0.216)	(0.306)	(0.220)	(0.220)
$RDev_{ijt}$		−0.331***	−1.199***		
		(0.037)	(0.142)		
$RDev_{ijt} \times ADisp_{it}$			0.769***		
			(0.121)		
$RRating_{ij}$				0.295***	−0.070
				(0.026)	(0.138)
$RRating_{ij} \times AMRating_{it}$					0.091***
					(0.034)
W_{ijm}	0.0156	0.0098	0.0280	0.0081	0.0053
	(0.1257)	(0.1245)	(0.1312)	(0.1274)	(0.1277)
Constant	3.173***	3.595***	5.106***	1.786**	2.890***
	(0.718)	(0.720)	(0.767)	(0.740)	(0.848)
Observations	23,368	23,368	23,368	23,368	23,368
Log likelihood	−16,294.2	−16,251.8	−16,230.7	−16,227.0	−16,223.3

Notes. Standard errors are in parentheses. Fixed effects at the category level and random effects at the app level are included. Ln indicates the variable is log transformed.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

most helpful favorable review along with the most helpful critical review prominently, before showing the most recent reviews. Furthermore, review platforms can use an analysis of the review text (length, tone, cognitive mechanisms, etc.) to determine which reviews consumers may find helpful, instead of relying solely on the helpfulness score. For instance, our results indicate that consumers find longer reviews more

helpful, perhaps because they appreciate the details contained in such reviews. In summary, understanding the rationale that underlies positive–negative asymmetry provides additional ways to sort, emphasize, and highlight those reviews that consumers may find useful but that may not have received high helpfulness scores.

Our study also has a few limitations that provide avenues for future research. First, our empirical approach cannot effectively uncover the exact reasons underlying the observed confirmation bias or reveal whether consumers are aware of this bias as they evaluate reviews. Furthermore, observational data cannot reveal how consumers cognitively process the information in reviews, and laboratory experiments could be an alternative method to answer these questions and extend the findings. Second, summary statistics of ratings are not the only source of information for consumers to form initial beliefs about products before they read and evaluate reviews. Future research can explore other sources of consumers' initial beliefs, such as social media recommendations, that are not easy to quantify with our data set. Third, there are many unobserved factors affecting consumers' perceived helpfulness of a specific review that we cannot control for in our analysis. Fourth, our data sample is from Apple's app market, so the generalizability of our findings may be limited to similar digital products. Future studies may want to sample a larger set of products to test whether our results can still hold in more general contexts.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/isre.2015.0617.

Acknowledgments

The authors thank the senior editor, the associate editor, and three anonymous reviewers for their constructive guidance during the review process. The authors are grateful to Andrew Burton-Jones, Katherine Stewart, and Lin Jiang for their insightful feedback on earlier versions of this paper.

References

Alba JW, Broniarczyk SM, Shimp TA, Urbany JE (1994) The influence of prior beliefs, frequency cues, and magnitude cues on consumers' perceptions of comparative price data. *J. Consumer Res.* 21(2):219–235.

Basuroy S, Chatterjee S, Ravid SA (2003) How critical are critical reviews? The box office effects of film critics, star power, and budgets. *J. Marketing* 67(4):103–117.

Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD (2001) Bad is stronger than good. *Rev. General Psych.* 5(4):323–370.

Blackwell M, Iacus SM, King G, Porro G (2009) CEM: Coarsened exact matching in Stata. *Stata J.* 9(4):524–546.

Cao Q, Duan W, Gan Q (2011) Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems* 50(2):511–521.

Cheung MY, Luo C, Sia CL, Chen H (2009) Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. *Internat. J. Electronic Commerce* 13(4):9–38.

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

Clemons EK, Gao G, Hitt LM (2006) When online reviews meet hyperdifferentiation: A study of the craft beer industry. *J. Management Inform. Systems* 23(2):149–171.

Darley JM, Gross PH (1983) A hypothesis-confirming bias in labeling effects. *J. Personality Soc. Psych.* 44(1):20–33.

Duan W, Gu B, Whinston AB (2008) The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry. *J. Retailing* 84(2):233–242.

Festinger L (1962) *A Theory of Cognitive Dissonance* (Stanford University Press, Stanford, CA).

Fischer P, Fischer J, Weisweiler S, Frey D (2010) Selective exposure to information: How different modes of decision making affect subsequent confirmatory information processing. *British J. Soc. Psych.* 49(4):871–881.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* 19(3):291–313.

Ghose A, Han SP (2014) Estimating demand for mobile applications in the new economy. *Management Sci.* 60(6):1470–1488.

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.* 23(10):1498–1512.

Grilli L, Rampichini C (2007) A multilevel multinomial logit model for the analysis of graduates' skills. *Statist. Methods Appl.* 16(3): 381–393.

Gunning R (1968) *The Technique of Clear Writing* (McGraw-Hill, New York).

Hart W, Albarracín D, Eagly AH, Brechan I, Lindberg MJ, Merrill L (2009) Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psych. Bull.* 135(4):555–588.

He SX, Bond SD (2015) Why is the crowd divided? Attribution for dispersion in online word of mouth. *J. Consumer Res.* 41(6): 1509–1527.

Helson H (1964) *Adaptation-Level Theory* (Harper, New York).

Iacus SM, King G, Porro G (2012) Causal inference without balance checking: Coarsened exact matching. *Political Anal.* 20(1):1–24.

Klayman J, Ha YW (1987) Confirmation, disconfirmation, and information in hypothesis testing. *Psych. Rev.* 94(2):211–228.

Korfiatis N, García-Bariocanal E, Sánchez-Alonso S (2012) Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Res. Appl.* 11(3):205–217.

Moe WW, Trusov M (2011) The value of social dynamics in online product ratings forums. *J. Marketing Res.* 48(3):444–456.

Mudambi SM, Schuff D (2010) What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quart.* 34(1):185–200.

Nickerson RS (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. General Psych.* 2(2):175–220.

Pan Y, Zhang JQ (2011) Born unequal: A study of the helpfulness of user-generated product reviews. *J. Retailing* 87(4):598–612.

Park J, Konana P, Gu B, Kumar A, Raghunathan R (2013) Information valuation and confirmation bias in virtual communities: Evidence from stock message boards. *Inform. Systems Res.* 24(4):1050–1067.

Pennebaker JW, Booth RJ, Francis ME (2007) Linguistic inquiry and word count (LIWC2007): A computer-based text analysis program. LIWC.net, Austin, TX.

Petrocelli JV, Tormala ZL, Rucker DD (2007) Unpacking attitude certainty: Attitude clarity and attitude correctness. *J. Personality Soc. Psych.* 92(1):30–41.

Petty RE, Briñol P, Tormala ZL, Wegener DT (2007) The role of meta-cognition in social judgment. Higgins ET, Kruglanski AW, eds. *Social Psychology: Handbook of Basic Principles* (Guilford Press, New York), 254–284.

Qiu L, Pang J, Lim KH (2012) Effects of conflicting aggregated rating on eWOM review credibility and diagnosticity: The moderating

role of review valence. *Decision Support Systems* 54(1):631–643.
Rabe-Hesketh S, Skrondal A, Pickles A (2002a) Multilevel selection models using GLLAMM. *Stata User Group Meeting, Maastricht, Netherlands*.
Rabe-Hesketh S, Skrondal A, Pickles A (2002b) Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata J.* 2(1):1–21.
Risen J, Gilovich T (2007) Informal logical fallacies. Sternberg RJ, Halpern D, Roediger H, eds. *Critical Thinking in Psychology* (Cambridge University Press, New York), 110–130.
Rucker DD, Tormala ZL, Petty RE, Briñol P (2014) Consumer conviction and commitment: An appraisal-based framework for attitude certainty. *J. Consumer Psych.* 24(1):119–136.
Scholz MPD, Dorner VD (2013) The recipe for the perfect review? *Bus. Inform. Systems Engrg.* 5(3):141–151.
Sen S, Lerman D (2007) Why are you telling me this? An examination into negative consumer reviews on the Web. *J. Interactive Marketing* 21(4):76–94.
Sherif M, Sherif CW (1967) Attitude as the individual's own categories: The social judgment-involvement approach to attitude and attitude change. Sherif CW, Sherif M, eds. *Attitude, Ego-Involvement, and Change* (Wiley, New York), 105–139.
Smith RE, Swinyard WR (1988) Cognitive response to advertising and trial: Belief strength, belief confidence and product curiosity. *J. Advertising* 17(3):3–14.
Sun M (2012) How does the variance of product ratings matter? *Management Sci.* 58(4):696–707.
Swann WB, Griffin JJ, Predmore SC, Gaines B (1987) The cognitive-affective crossfire: When self-consistency confronts self-enhancement. *J. Personality Soc. Psych.* 52(5):881–889.
Trope Y, Bassok M (1982) Confirmatory and diagnosing strategies in social information gathering. *J. Personality Soc. Psych.* 43(1):22–34.
Vaish A, Grossmann T, Woodward A (2008) Not all emotions are created equal: The negativity bias in social-emotional development. *Psych. Bull.* 134(3):383–403.
Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).
Yin D, Bond SD, Zhang H (2014) Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quart.* 38(2):539–560.
Zhang JQ, Craciun G, Shin D (2010) When does electronic word-of-mouth matter? A study of consumer product reviews. *J. Bus. Res.* 63(12):1336–1341.