Swayed by the Reviews:

Disentangling the Effects of Average Ratings and Individual Reviews in Online Word-of-Mouth

Zhanfei Lei
Isenberg School of Management
University of Massachusetts Amherst
Amherst, Massachusetts 01003, USA
zlei@isenberg.umass.edu

Dezhi Yin
Muma College of Business
University of South Florida
Tampa, Florida 33620, USA
dezhiyin@usf.edu

Sabyasachi Mitra*
Warrington College of Business
University of Florida
Gainesville, Florida 32611, USA
Saby.Mitra@warrington.ufl.edu

Han Zhang
Scheller College of Business
Georgia Institute of Technology
Atlanta, Georgia 30308, USA
han.zhang@scheller.gatech.edu

*Correspondence author

***** *Forthcoming at Production and Operations Management* *****

Swayed by the Reviews:
Disentangling the Effects of Average Ratings and Individual Reviews in Online Word-of-Mouth

ABSTRACT

Online word-of-mouth studies generally assume that a product's average rating is the primary force shaping consumers' purchase decisions and driving sales. Similarly, practitioners place more emphasis on average ratings by displaying them at more salient places than individual reviews. In contrast, emerging evidence suggests that individual reviews also affect the decision-making of those consumers who consult both kinds of information. However, because average ratings and individual reviews are often correlated and confounded empirically, little research has attempted to disentangle their effects. To address this empirical challenge, we construct trade-off situations in which the average ratings and top-ranked reviews of different product options do not align with each other. We then investigate consumers' preferences that can indirectly reveal the relative impact of average ratings vs. top reviews. Through an archival analysis of a panel dataset and two laboratory experiments, we find consistent evidence for a swaying effect of individual reviews and reveal their textual content as a likely reason. These findings challenge the commonly accepted assumption of average ratings being the primary driver of consumers' purchase decisions and suggest that consumers may not be as rational as previous literature assumed. In addition, this paper is the first to disentangle the effects of average ratings and individual reviews on consumer decision-making and to explore a possible reason for the swaying effect of individual reviews. Our paper illustrates the importance of information accessibility in consumers' purchase decisions, and our findings offer valuable insights for product manufacturers, online retailers, and review platforms.

***Key words**:* average ratings; individual reviews; information accessibility; consumer decision-making; online word-of-mouth

History: Received: July 2021; Accepted: January 2022 by Subodha Kumar, after one revision.

INTRODUCTION

Online reviews are valuable and influential in consumers' purchase decisions (e.g., Chevalier and Mayzlin, 2006; Forman et al., 2008). A recent survey conducted in April 2021 asked consumers about the factors that impact their online purchase decisions (PowerReviews, 2021). Based on over 6,500 responses across the U.S., 94% of the consumers indicated online reviews as the most important factor, followed by product price (91%), free shipping (78%), brand (65%), and friend/family recommendations (60%). Given this trend, businesses have incorporated consumer reviews into their marketing strategies and adjusted services based on the opinions expressed in reviews. In addition, e-commerce and review platforms constantly tweak the design and operation of review systems to gain a strategic advantage (Gutt et al., 2019).

The role of online reviews has also attracted considerable attention and interest from researchers (for recent reviews, see Gutt et al., 2019; Jabr et al., 2020). One stream of research investigated reviewers' biases and information updating processes when they write reviews (Chen et al., 2016; Sun and Xu, 2018; Wang et al., 2019), as well as factors that influence seller strategies to manipulate consumer reviews (e.g., Guan et al., 2020; Shen et al., 2015). The second stream of research examined the impact of online reviews (and management responses to reviews) on consumer perceptions (e.g., Qahri-Saremi and Montazemi, 2019) and satisfaction (e.g., Gu and Ye, 2014; Yan et al., 2019). The third stream of research investigated the influence of online reviews on product sales and consumer purchase decisions (Babić Rosario et al., 2016; Floyd et al., 2014), such as exploring the dynamics between online reviews and sales (e.g., Ceran et al., 2016; Duan et al., 2008b), designing review-based big data methodologies to improve sales forecasting (e.g., Lau et al., 2018), and studying how characteristics of review attributes, products, and consumers affect the role of online reviews in product sales (e.g., Ba et al., 2020; Liu and Karahanna, 2017). For a literature review of related papers in operation management, see Appendix A in E-Companion.

With a focus on extending the third stream of the research described above, this work aims to explore a more nuanced role of online reviews by disentangling the effects of online ratings and individual reviews on consumer decision-making. Review platforms and many online retailers allow consumers to share their opinions of a product in text reviews along with ratings (typically 1 to 5 stars). To further help prospective consumers easily gauge collective opinions, review platforms and retailers universally display the products' average ratings in the most prominent places, such as the front page, product listing pages, and search result pages. These prominently displayed average ratings are critical in the online shopping process because they are believed to reflect product quality (De Langhe et al., 2015) and are the basis of consumers' initial impressions about different product options (Yin et al., 2016). Such initial impressions help consumers simplify their purchase decisions by narrowing down the number of

product options to a smaller set, namely a consideration set—a subset of available options to which consumers limit their attention and evaluation (Roberts and Lattin, 1991; Wright and Barbour, 1977). To choose among options in the consideration set, consumers may next seek and read more information such as some individual reviews of each product. Among consumers who are seriously considering a purchase, the majority consult both average ratings and some reviews to make purchase decisions (Liu et al., 2019).[1] A natural question is: which input dominates consumers' final purchase decisions?

It is commonly assumed by researchers and practitioners that a product's average rating should play a greater role than individual reviews in consumers' purchase decisions and product sales. Prior empirical studies that examined the driving forces behind product sales focused primarily on the impact of average ratings and other summary rating statistics (see Babić Rosario et al., 2016; Floyd et al., 2014), while few studies considered the influence of both aggregated ratings and individual reviews (as a few exceptions, see Jabr and Rahman, forthcoming; Liu et al., 2019; Vana and Lambrecht, 2021). This prevailing focus on the average product rating is not surprising because it incorporates all the historical evaluations from prior reviewers, is routinely used by consumers to infer product quality (De Langhe et al., 2015), and is more representative than individual reviews (Liu and Karahanna, 2017). In practice, review sites and online retailers also display the average product rating at more prominent places than individual reviews, such as at the top of product review pages and product listing pages.

Although a product's average rating is all-encompassing and a comprehensive measure of product quality, most consumers who have high purchase intention also consult individual reviews before making up their minds. A growing literature studying individual reviews has examined the influence of rating and review text characteristics on the helpfulness of reviews (e.g., Lei et al., 2021; Mudambi and Schuff, 2010; Wang et al., 2019; Yin et al., 2017; Yin et al., 2016) because separating helpful from unhelpful reviews reduces information overload and improves efficiency for consumers. In addition, the latest empirical and experimental studies revealed that a small set of top-ranked individual reviews could also affect consumers' attitudes and purchase decisions (e.g., Liu et al., 2019; Yin et al., 2021). Thus, Watson et al. (2018) called for more research to explore how consumers integrate summary rating statistics and individual reviews in their decision-making process.

While a few recent papers have started to tackle this challenge (Jabr and Rahman, forthcoming; Liu et al., 2019; Vana and Lambrecht, 2021), they revealed only the effect of individual reviews *above and*

[1] We acknowledge that other information such as price, brand image, and the total number of ratings could also be important inputs in consumer purchase decisions. However, we decided to focus on the average rating because a) the average rating is among the most important factors revealed to influence product sales (Floyd et al., 2014), b) the quality of a product reflected by the average rating is relatively time-invariant and independent from the company's advertising efforts, and c) the product's average rating and the average valence of top reviews are directly comparable.

beyond average ratings, and they relied only on secondary data (see Table 1). On the other hand, no research to our knowledge has *disentangled* the effects of average ratings and a few top-ranked reviews on consumers' purchase decisions. In the current paper, we take on this challenge through a trade-off paradigm that is capable of comparing the relative weights or importance of different attributes in consumer decision-making (Ajzen, 2008).[2] Specifically, we conduct an archival analysis of secondary data and two experiments to examine consumers' purchase preferences between product options whose average ratings do not align with the valence of top-ranked reviews. Such a trade-off design allows us to disentangle the effects of average ratings and individual reviews that are often confounded in empirical analyses of secondary data (e.g., Liu et al., 2019). The trade-off design has also been used to disentangle the effects of different summary rating statistics (see Watson et al., 2018). In addition, the trade-off between average ratings and top reviews is not uncommon in reality because reviewers may provide ratings (that are incorporated into average ratings) without leaving a review in most review platforms and because the top-ranked reviews may not be as representative as the average ratings. Through the three studies, we not only find evidence supporting a swaying effect—a greater persuasive impact—of top-ranked reviews, but also explore a possible mechanism behind this effect.

Table 1: Related Literature Examining the Impacts of Average Ratings and Individual Reviews

Research	**Objectives**		**Dependent Variables**		**Methods**	
	Explore the effect of individual reviews *above and beyond* average ratings by including both in regressions	***Disentangle the effects* of individual reviews and average ratings through a trade-off design**	**Product sales**	**Purchase likelihood**	**Empirical analysis**	**Lab experiment**
Liu et al. (2019)	√			√	√	
Jabr and Rahman (forthcoming)	√		√		√	
Vana and Lambrecht (2021)	√			√	√	
This Research		√	√	√	√	√

This work offers three notable contributions to the online word-of-mouth literature. First and foremost, our research challenges the conventional wisdom that a product's average rating and other

[2] It is worth pointing out that not all consumers choose to read reviews, and only consumers who are seriously considering the purchase of a product included in their consideration sets would consult its reviews (Liu et al., 2019). We limit our focus to the purchase behavior of those consumers who consult both average ratings and individual reviews because such consumers are the primary targets of retailers and product manufacturers, and also because it is impossible to disentangle the effects of average ratings and individual reviews when consumers do not read any reviews.

summary rating statistics are the primary determinants of consumers' purchase decisions (Babić Rosario et al., 2016; Floyd et al., 2014). Since the average product rating is the most comprehensive signal of product quality available to consumers (De Langhe et al., 2015), it is reasonable to assume that a rational consumer's decision-making process is heavily influenced by this prominent and all-encompassing signal of product quality. However, the majority of consumers who are serious about purchases also read individual reviews before making up their minds (Liu et al., 2019). Our demonstration of the swaying effect of individual reviews suggests that consumers may not be as "rational" as the online word-of-mouth literature assumed and that consumers' purchase decisions can be biased towards a few top-ranked reviews. Moreover, because the assumption of a product's average rating being the primary driver of purchase decisions may not apply to consumers who integrate both reviews and average ratings in their decision-making, prior literature (especially the third stream we reviewed earlier) may have exaggerated the effect of a product's average rating and other summary rating statistics.

Second, this paper represents an initial attempt to disentangle the effects of the average rating and individual reviews for consumers who consult and integrate both types of information in their final decisions. Recent empirical studies have examined the effect of individual reviews *above and beyond* average ratings (Jabr and Rahman, forthcoming; Liu et al., 2019; Vana and Lambrecht, 2021). While these studies provided valuable insights regarding the importance of individual reviews, we are not aware of any studies that attempt to disentangle the effects of average ratings and individual reviews that are often confounded in empirical studies. Through a trade-off design where the "stars" of two distinct sources do not align with each other, our findings from one archival and two experimental studies provide compelling evidence for a swaying effect of individual reviews, suggesting that a few top reviews may play a more dominant role than the average rating in consumer purchase decisions. Although average product ratings might still be an important factor turning consumers away (Liu et al., 2019), our findings indicate that the greater persuasive power of a few top-ranked reviews should not be overlooked.

Third, our results reveal a possible underlying mechanism for the swaying effect, deepening our understanding of why a few top-ranked reviews may shift consumers' preferences between multiple choices. Specifically, this research highlights the critical role of information accessibility in online word-of-mouth. Building on the mere-accessibility framework, ease-of-retrieval explanation, and cognitive elaboration arguments (Kisielius and Sternthal, 1986; Mafael et al., 2016; Schwarz et al., 1991), the results of our final experiment reveal that the swaying effect of individual reviews is likely driven by the reviews' textual content rather than by their ratings. The mere-accessibility framework suggests that decision-makers primarily rely on accessible information that can be brought to mind and accessed from memory easily when they make a decision (Menon and Raghubir, 2003). In our context, the review content is more concrete and easier to recall than the review rating. Ultimately, the concrete details

conveyed in the review text drive the swaying effect of individual reviews. Revealing the importance of information accessibility in consumers' purchase decisions opens up opportunities for future research.

HYPOTHESES AND THEORY DEVELOPMENT

Average Product Ratings

Because little research has examined the impact of individual reviews (except a few most recent empirical explorations in Liu et al., 2019; Vana and Lambrecht, 2021), large-scale meta-analyses of empirical evidence on determinants of product sales focus entirely on the role of average product ratings and other summary rating statistics (see Babić Rosario et al., 2016; Floyd et al., 2014). Although the average rating may not necessarily reflect the product's true quality (e.g., Duan et al., 2008a; Hu et al., 2009), consumers generally believe them to be strongly associated (De Langhe et al., 2015). In addition, because a product's average rating aggregates all the historical opinions from prior customers (Sun, 2012), a rational consumer should rely more on the all-encompassing average rating than a small set of individual reviews in making purchase decisions. Accordingly, researchers typically assume the average ratings to be the primary input in consumer decision-making and a primary driver of product sales.

In practice, nearly all review sites present the average rating more prominently and frequently than individual reviews. Because of its intuitive association with product quality, the average rating is assumed to be among the most essential pieces of information for consumers. Thus, review sites typically display average ratings at very salient places: right beside product options returned after a search and at the top of product pages. For example, when consumers search on Amazon.com, the resulting page includes a list of relevant product options, each accompanied by the average rating displayed on a 5-star rating scale, as well as other information (such as a picture and a price). Consumers can click on a product of interest and see its average rating again before they can scroll to the bottom of the page to read individual reviews.

This commonly adopted display strategy also conforms to the saliency effect and the observational learning effect of product-level signals. Specifically, the saliency effect refers to the phenomenon that environmental signals which receive more attention (i.e., are more salient) are likely to be weighted more heavily in subsequent judgments and decisions (Taylor and Thompson, 1982). Similarly, the observational learning effect refers to the phenomenon that people's decisions are influenced by observing others' actions that provide helpful information (Banerjee, 1992; Bikhchandani et al., 1992). In online word-of-mouth, product-level signals that are both salient and reflect prior consumers' judgments have been shown to be powerful predictors of consumer decisions (Cai et al., 2009; Salganik et al., 2006). Because the overall evaluation of all prior customers of a product is encapsulated in its average rating, and due to its salient display on nearly all review platforms, it is reasonable to expect the average product rating to be more influential than individual reviews in a rational consumer's purchase decisions. Following the preceding reasoning, we propose the first hypothesis below.

Hypothesis 1a: Given a choice set of product options, consumers' purchase decisions are influenced more by the average ratings than the individual reviews.

It is important to note that the effects of average ratings and individual reviews can only be compared when both are assessed and read by consumers. Otherwise, disentangling the effects of these two information cues would be impossible. For instance, before consumers consult any individual reviews, average ratings should be the primary cue for consumer decision-making because individual reviews would not have received any exposure yet. Such cases are out of this paper's scope.

Individual Reviews

Despite the importance of a product's average rating, most consumers who are serious about purchase would also seek and read a few individual reviews before making up their minds (Liu et al., 2019). Several existing studies have demonstrated that individual reviews can also influence consumers' purchase decisions and product sales (e.g., Liu et al., 2019; Vana and Lambrecht, 2021; Yin et al., 2021). However, we are not aware of any research that disentangles the effects of average ratings and top-ranked reviews that are likely to be read by most consumers.

Drawing on the mere-accessibility framework, we argue that the average product rating is not necessarily the most important predictor of consumer purchase decisions. The mere-accessibility framework states that people rely primarily on accessible information in their decision-making through an unintentional and effortless process (Menon and Raghubir, 2003). *Accessibility* of information refers to the ease with which the information can be brought to mind and accessed from memory when one makes a decision (Kisielius and Sternthal, 1986). Decision-makers often associate the probability of a target object with its accessibility because of their natural co-occurrences; for example, more frequent examples are recalled better and faster than less frequent examples, and more likely instances are easier to imagine than unlikely instances (termed "availability heuristic"; see Tversky and Kahneman, 1973). People often use the ease with which information can be brought to mind to infer the frequency and importance of such information (Schwarz et al., 1991). As a result, decision-makers use easy-to-retrieve and accessible instances as the primary input to their judgment and decisions (commonly called the ease-of-retrieval effect).

Among a variety of factors that contribute to information accessibility, the concreteness of information serves as one of the most critical contributors (Nisbett and Ross, 1980).[3] Compared with abstract information, concrete information is easier to imagine and thus easier to retrieve and access from memory. In our context, individual reviews could be more influential than average ratings in consumers'

[3] In addition to concreteness, the salience (i.e., received attention) of stimuli could be another contributor of information accessibility (Higgins, 1996). The average rating is the all-encompassing and most comprehensive signal of a product's quality, and its visual format should also enhance its salience above any individual review. This argument would support H1a (as explained earlier) and present a counterargument for H1b.

purchase decisions because, unlike a product's average rating, individual reviews contain detailed, concrete experiences and opinions. When consumers make purchase decisions, concrete information in individual reviews (as opposed to abstract average ratings) could be easier to recall from memory and more accessible, thus being perceived as important and typical of a possible consumption experience with the product. According to the mere-accessibility framework and ease-of-retrieval explanation, consumers' purchase decisions might be influenced more by easy-to-recall individual reviews than the average product rating, and we propose a competing hypothesis.

Hypothesis 1b: Given a choice set of product options, consumers' purchase decisions are influenced more by the individual reviews than the average ratings.

To examine the competing hypotheses, we adopted a trade-off design with two (groups of) options whose average ratings contradict top reviews. Specifically, one option (group) is superior based on average ratings, and the other option (group) is superior based on individual reviews. Such a trade-off design is derived from a common paradigm in experimental research on the relative weights or importance of different attributes (defined as relevant factors that differentiate between alternatives) in consumers' decision-making processes (Ajzen, 2008). Within this paradigm, studies typically present two alternatives (that vary the two attributes simultaneously and involve a tradeoff between the two attributes) and then ask participants to decide which alternative they would purchase based on all the provided information. Their revealed preferences can indirectly indicate which of the two attributes is more important and weighted more heavily (e.g., Shiv and Fedorikhin, 1999).

We conducted one archival and two experimental studies with such a trade-off design. In Study 1, we utilized a unique panel dataset collected daily from Apple's App Store over a two-month period and compared the download rankings of two groups of apps (through a trade-off design) to empirically test the relative impacts of the average rating and top reviews. In Study 2, we replicated the findings of Study 1 through an experiment in which participants were presented with two choice options involving a trade-off between average ratings and top reviews. In Study 3, we explored the source of the effect identified in the first two studies, differentiated the role of the rating and textual content of individual reviews, and ruled out an alternative explanation.

STUDY 1

The primary goal of this initial study was to test the competing hypotheses in a real-world setting with actual ratings and reviews of apps from Apple's App Store. Existing users of an app can evaluate the app by assigning a rating on a scale of 1 to 5 stars. In addition, users can provide a detailed description of their experiences with the app in a text review. Users can also submit a rating without writing a text review. When prospective consumers read the text reviews of an app, they can indicate whether they find a review "helpful" or "not helpful" by clicking on the "Yes" or "No" buttons next to the review. By

default, at the time of data collection, the review section of an app displayed 10 reviews per page, sorted by review helpfulness.

Data and Variables

We collected daily reviews of apps from Apple's App Store for a period of two months (62 consecutive days). We targeted 538 apps ranked in the top 100 in Apple's App Store at least once during the month prior to our data collection. Apple classified all apps into 21 categories (such as games, business, finance, and news). The App Store organized the reviews of each app into 10 reviews per page, and the 10 reviews on the first page are the most observable for prospective consumers. Apple displayed the 10 most helpful reviews on the first page by default, and these reviews may change on different days as new votes are cast and review helpfulness scores are updated. It is reasonable to expect that most consumers did not change the sorting order of the reviews and saw the same 10 first-page reviews of an app on the same day. Accordingly, we did not change the default sorting order of reviews in our data collection, and we extracted the rating and content of each app's 10 first-page reviews daily.

For each app, we also tracked the following app-level data that changed over time: the overall ranking of the app, its average rating, the total number of ratings, the distribution of the ratings (e.g., number of one-star ratings, number of two-star ratings, etc.), the price of the app, whether the app released an updated version on a specific date, and the number of days since the app was first launched. It should be noted that a user often provides a rating for the app but does not write a text review. Thus, an app typically has many more ratings (based on which the app's average rating is calculated) than reviews. Our final sample contained 482 (out of 538) apps that had the information needed to calculate all the variables in our model during the study period.

Table 2 shows the definitions of the variables used in the empirical analysis, while Tables 3 and 4 show statistics and correlations for these variables. For the variable definitions in Table 2, *i* indexes an app and *t* indexes the event time (day) during our study period. Our dependent variable is $ARank_{it}$, the overall rank of app *i* at time *t* (log transformed). It is calculated based primarily on the number of app downloads and is displayed in the "Top Charts" list made available by Apple. We collected the overall rank of each app in our sample for each day in the study period from Apple's App Store. The overall rank ($ARank_{it}$) is a proxy for product sales in the online app context. A smaller numeric rank indicates a greater number of downloads.

Table 2: Variable Definitions

Variable Name	Operationalization
$ARank_{it}$	Rank of app *i* at time *t* based on the number of downloads
$RARating10_{it}$	Average rating of the first 10 reviews of app *i* at time *t*
$ARating_{it}$	Average rating of app *i* at time *t (based on all ratings for the app)*

$ARCount_{it}$	Cumulative number of ratings for app *i* at time *t*
$APrice_{it}$	Price per download of app *i* at time *t*
$AUpd_{it}$	= 1 if app *i* released an update (new version) at time *t*, 0 otherwise
$ADispersion_{it}$	Standard deviation of the ratings for app *i* at time *t*
$ADays_{it}$	Age of app *i* (in days) at time *t*
$RALength10_{it}$	Average number of words in the first 10 reviews of app *i* at time *t*

Table 3: Descriptive Statistics

Variable	N	Mean	Std Dev	Min	Max
$ARank_{it}$	23359	346.44	399.77	1	1500
$RARating10_{it}$	30094	3.83	1.00	1	5
$ARating_{it}$	29425	4.18	0.62	1	5
$ARCount_{it}$	30094	2780.94	7621.30	0	104407
$APrice_{it}$	30094	1.25	4.37	0	69.99
$AUpd_{it}$	30094	0.02	0.14	0	1
$ADispersion_{it}$	29425	1.14	0.35	0	2
$ADays_{it}$	30094	538.80	506.14	3	1877
$RALength10_{it}$	30094	26.34	16.22	0.25	235.5

Table 4: Pairwise Correlations

	Variable	1	2	3	4	5	6	7	8	9
1	$ARank_{it}$	1.00								
2	$RARating10_{it}$	0.03*	1.00							
3	$ARating_{it}$	-0.02*	0.58*	1.00						
4	$ARCount_{it}$	-0.15*	-0.12*	0.17*	1.00					
5	$APrice_{it}$	0.03	0.10*	0.13*	-0.05*	1.00				
6	$AUpd_{it}$	-0.01*	0.01	-0.03*	-0.04*	-0.02*	1.00			
7	$ADispersion_{it}$	0.00	-0.54*	-0.79*	-0.18*	-0.10*	0.02*	1.00		
8	$ADays_{it}$	-0.25*	-0.07*	-0.07*	0.04*	0.02*	-0.01	0.04*	1.00	
9	$RALength10_{it}$	-0.01	-0.33*	-0.19*	0.04*	0.23*	-0.03*	0.18*	0.11*	1.00

Pairwise correlations shown in the table. * $p < 0.05$

The other variables in our analysis are defined as follows. $RARating10_{it}$ is the average rating of the 10 most observable reviews (that appear on the first page of reviews) for app *i* at time *t*. It is important to note that consumers could see the actual text along with a rating for any particular review, and that $RARating10_{it}$ is simply a proxy for the valence of the top reviews that consumers see; its value is not actually displayed on the page. $ARating_{it}$ is the overall average rating of app *i* at time *t* based on all the ratings provided by consumers for the latest version of the app. The App Store site displays the average rating of each app prominently along with other summarized statistics (such as the number of ratings), so

we obtained $ARating_{it}$ directly from the App Store. We also define a derived variable $Diff10_{it}$ as the difference between the average rating of the 10 most observable reviews on the first review page and the average rating of the app ($Diff10_{it} = RARating10_{it} - ARating_{it}$). $ARCount_{it}$ is the cumulative number of ratings for app *i* at time *t*. The number of ratings of an app can affect consumers' purchase decisions since it indicates the popularity of the app. $ADispersion_{it}$ is the dispersion of ratings measured by the population standard deviation of all the ratings for app *i* at time *t*. The dispersion of a product's ratings can shape consumers' confidence in their first impressions of the product (Yin et al., 2016). $APrice_{it}$ is the price of app *i* at time *t*. The price of an app may change over time due to promotions or other reasons, and the price could affect consumers' download decisions. $AUpd_{it}$ is an indicator which equals 1 if app *i* released an update (new version) at time *t*, and 0 otherwise. Consumers' intention to purchase or download an app can be influenced by whether the app has an updated version, because a new version often contains improvements and there may be promotions associated with a new version. $ADays_{it}$ indicates the number of days at time *t* since the app was first launched. Finally, $RALength10_{it}$ is the average number of words in the 10 reviews on the first review page of app *i* at time *t*.

Methods and Empirical Analysis

Before evaluating our competing hypothesis, we attempted to replicate the recent empirical findings that top reviews can influence sales above and beyond average ratings under certain conditions (Jabr and Rahman, forthcoming; Liu et al., 2019; Vana and Lambrecht, 2021). We performed an analysis to evaluate how the difference between the ratings of the 10 reviews on the first review page and the average rating of the app ($Diff10_{it}$) affects app rankings. Specifically, we evaluated the following in STATA.

$$Ln(ARank_{it+1}) = \beta_0 + \beta_1 ARCount_{it} + \beta_2 APrice_{it} + \beta_3 AUpd_{it} + \beta_4 ADispersion_{it} + \beta_5 ADays_{it} + \beta_6 RALength10_{it} + \beta_7 ARating_{it} + \beta_8 Diff10_{it} + U_i + \in_{it} \quad (1)$$

In (1), U_i is the app level fixed effects and $\in_{it}$ is the error term. The results are shown in Table 5. Columns (1) and (2) show the results with the full sample. In column (2), the coefficient of the $ARating_{it}$ variable was significant and negative (β = -0.103, $p < 0.001$), as expected, indicating that app ranking improved (download increased) as the average rating of the app increased. The coefficient of the $Diff10_{it}$ variable was significant and negative (β = -0.032, $p < 0.001$), indicating that as the ratings of the 10 reviews on the first page deviated more positively from the average rating of the app, the app rank improved (downloads increased). Thus, differences of the most observable reviews from the average rating of the app influenced app downloads, replicating previous empirical findings.

To explore the conditions under which top reviews may influence app sales beyond the impact of average ratings, we conducted a split sample analysis based on a cutoff threshold of 3 stars for the average ratings because 3 is the middle point of the 5-point rating scale. Results of this analysis shown in columns (3) and (4) illustrate an interesting observation. The coefficient of $Diff10_{it}$ was significant and

negative in column (3) but not in column (4), indicating that reviews mattered only when the average rating of the app was above a certain threshold (3.0) that put the app in the consideration set of the consumer. At or below the threshold, consumers probably did not read the reviews as the app may not be in their consideration set. To verify the correctness of this threshold, we tried out different cutoff values (2.5, 3.0, 3.5 and 4.0) in columns (4) through (7) when the sample was restricted to the reviews of apps whose average ratings were below the specific cutoff. We found that the coefficient for $Diff10_{it}$ was negative and significant for cutoff thresholds higher than 3.0. However, the coefficient was not significant when the cutoff threshold was 3.0 or lower, suggesting that 3.0 was a likely threshold. Consumers in our dataset probably did not read reviews for apps whose average rating was below 3.0. Accordingly, to test the proposed competing hypotheses (see later), we focused on apps with an average rating of at least 3.5.

Table 5: Effect of Differences from Average Ratings

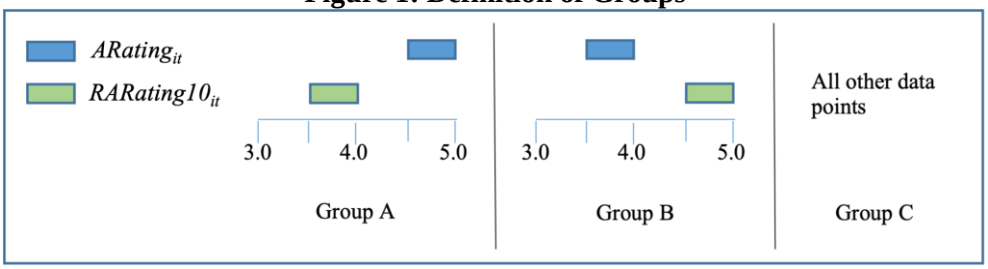
	Full Sample		$ARating_{it}$		$ARating_{it}$		
			> 3.0	≤ 3.0	≤ 2.5	≤ 3.5	≤ 4.0
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$ARCount_{it}$(ln)	-0.019***	-0.022***	-0.020***	-0.048**	-0.056*	-0.069***	-0.049***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.03)	(0.00)	(0.00)
$APrice_{it}$	0.320***	0.320***	0.324***	0.349	0.380	0.165***	0.237***
	(0.00)	(0.00)	(0.00)	(0.21)	(0.41)	(0.00)	(0.00)
$AUpd_{it}$	-0.119***	-0.122***	-0.094***	-0.269**	-0.253*	-0.170**	-0.117**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)
$ADispersion_{it}$	-0.244***	-0.259***	-0.153***	-0.548***	-0.336*	-0.806***	-0.768***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)
$ADays_{it}$(ln)	0.935***	0.929***	0.929***	0.579***	-0.619***	0.939***	1.084***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$RALength10_{it}$	0.003***	0.002***	0.002***	0.001	0.000	0.003***	0.001**
	(0.00)	(0.00)	(0.00)	(0.35)	(0.96)	(0.00)	(0.01)
$ARating_{it}$	-0.087***	-0.103***	-0.042	-0.157*	-0.162	-0.140***	-0.130***
	(0.00)	(0.00)	(0.14)	(0.01)	(0.12)	(0.00)	(0.00)
$Diff10_{it}$		-0.032***	-0.035***	0.007	0.017	-0.082***	-0.056***
		(0.00)	(0.00)	(0.77)	(0.53)	(0.00)	(0.00)
Intercept	-0.073	0.069	-0.311	2.788***	10.013***	1.307***	0.358
	(0.65)	(0.68)	(0.15)	(0.00)	(0.00)	(0.00)	(0.13)
Fixed Effects	App	App	App	App	App	App	App
N	22423	22423	20587	1836	957	4230	8598
R^2	0.12	0.12	0.13	0.04	0.09	0.11	0.15

p-values in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

To evaluate our competing hypotheses, we created two groups of data points with a tradeoff between the overall average rating ($ARating_{it}$) and the valence of the top 10 reviews ($RARating10_{it}$). Specifically, Group A contained 2902 data points where $ARating_{it} \geq 4.5$ and $RARating10_{it}$ was between 3.5 and 4.0. Likewise, Group B contained 1044 data points where $ARating_{it}$ was between 3.5 and 4.0 and $RARating10_{it} \geq 4.5$. Additionally, Group C contained the remaining 26,148 data points that

were not included in Groups A or B (see Figure 1). The construction of Group A and Group B follows the trade-off design commonly used to examine the relative importance of two attributes in consumers decision-making (Ajzen, 2008). In our case, data points in Group A have a higher value for $ARating_{it}$ than those in Group B, but data points in Group B have a higher value for $RARating10_{it}$ than those in Group A. If consumers' purchase decisions are influenced more by the average ratings than the individual reviews (as we proposed in H1a), data points in Group A should be associated with more app downloads (i.e., lower app ranks) than data points in Group B. On the contrary, if individual reviews play a greater role in consumers' purchase decisions (as we proposed in H1b), data points in Group B should be associated with more app downloads than those in Group A. The empirical approaches described below compare the impacts of data points in Groups A and B on the app ranking in the next period ($ARank_{it+1}$).

Figure 1: Definition of Groups



Our first empirical approach accounts for app level heterogeneity through app level fixed effects. It is important to note that while the 3,946 data points in Groups A and B were spread across 227 apps, only 30 of the apps had at least one data point in each group, and only 18 apps had more than one data point in each group. Consequently, app level fixed effects are infeasible when directly comparing Groups A and B. Instead, we first compared data points in Group A with all other data points (Groups B and C) and then compared data points in Group B with all other data points (Groups A and C) and finally tested for differences in the coefficients. We define a variable *GroupA* that is set to 1 if the data point belonged to Group A and 0 if it belonged to Groups B and C. Likewise, we define a variable *GroupB* that is set to 1 if the data point belonged to Group B and 0 if it belonged to Groups A and C. We evaluated the following empirical model in STATA where U_i is the app level fixed effect. Following prior literature (Chevalier and Mayzlin, 2006; Forman et al., 2008), we log transformed count variables and the dependent variable to account for scale effects (ranks of popular apps may change more) and ease interpretation (the coefficients approximately indicate a percentage change in rank).

$$Ln(ARank_{it+1}) = \beta_0 + \beta_1 ARCount_{it} + \beta_2 APrice_{it} + \beta_3 AUpd_{it} + \beta_4 ADispersion_{it} + \beta_5 ADays_{it} + \beta_6 RALength10_{it} + \beta_7 GroupA + \beta_8 GroupB + U_i + \in_{it} \quad (2)$$

The results of the estimation are shown in Table 6 (columns 1-4).[4] Column (1) includes the control variables. Time invariant and unobserved characteristics of an app are captured through the fixed effects intercept term in the model, and the coefficients of other variables capture within-app differences. As the number of ratings increased for an app, rankings improved (i.e., became numerically lower) in the next period (β = -0.02, $p < 0.001$). Higher rating dispersion of the app also improved the ranking (β = -0.156, $p < 0.001$). Price (β = 0.32, $p < 0.001$) and age of the app (β = 0.929, $p < 0.001$) had detrimental effects on the number of downloads, while releasing an updated version improved the ranking (β = -0.114, $p < 0.001$). Finally, shorter reviews were associated with improved rankings (β = 0.003, $p < 0.001$).

Table 6: Treatment versus Control Samples

	App Fixed Effects					Coarsened Exact Matching		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$ARCount_{it}$(ln)	-0.020***	-0.021***	-0.020***	-0.021***	-0.019***	-0.005	-0.007	-0.007
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.44)	(0.32)	(0.13)
$APrice_{it}$	0.320***	0.321***	0.321***	0.321***	0.318***	-0.009*	-0.012**	-0.008*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.03)	(0.01)	(0.02)
$AUpd_{it}$	-0.114***	-0.114***	-0.114***	-0.114***	-0.113***	-0.058	-0.050	0.000
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.40)	(0.46)	(1.00)
$ADispersion_{it}$	-0.156***	-0.152***	-0.146***	-0.142***	-0.137***	-0.004	0.117*	0.042
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.92)	(0.05)	(0.24)
$ADays_{it}$(ln)	0.929***	0.930***	0.929***	0.929***	0.935***	-0.066*	-0.070*	-0.037*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.01)	(0.04)
$RALength10_{it}$	0.003***	0.003***	0.003***	0.003***		0.001	0.001	
	(0.00)	(0.00)	(0.00)	(0.00)		(0.21)	(0.18)	
Group A		0.050**		0.049**	0.036*			
		(0.00)		(0.00)	(0.03)			
Group B			-0.079**	-0.077**	-0.041			
			(0.00)	(0.00)	(0.06)			
Group A vs. B							0.083**	0.046*
							(0.00)	(0.01)
$RALength5_{it}$					0.001***			0.001
					(0.00)			(0.08)
Intercept	-0.501***	-0.507***	-0.505***	-0.511***	-0.512***	7.358***	7.194***	7.327***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Fixed Effects	App	App	App	App	App	CEM Strata	CEM Strata	CEM Strata
N	22423	22423	22423	22423	22423	1440	1440	1850
R^2	0.12	0.12	0.12	0.12	0.12	0.96	0.96	0.97

p-values in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Column (2) introduces the *GroupA* variable. The coefficient was positive and significant (β = 0.05, $p < 0.01$) indicating that ranking worsened in the next period for data points in Group A (compared to all other data points) after accounting for app level heterogeneity. Likewise, column (3) introduces the

[4] While 23,359 data points had data on $ARank_{it}$, 936 of these 23,359 data points had missing values for one or more of the other control variables, leading to a final sample of 22,423 data points for this analysis.

GroupB variable. The coefficient was negative and significant (β = -0.079, $p < 0.01$) indicating that ranking improved in the next period for data points in Group B (compared to all other data points). Column (4) includes both the *GroupA* and *GroupB* variables. An F test rejected the hypothesis that the coefficients for *GroupA* and *GroupB* were equal (F(1, 21955) = 18.1, $p < 0.001$). Since Group A contains data points with higher average ratings while Group B contains data points with higher valence of the top 10 reviews, our results supported H1b. Further, since consumers may read fewer than 10 reviews on the page, column (5) in Table 6 repeats the analysis in column (4) with the *GroupA* and *GroupB* variables re-defined using the first 5 reviews on the page. We also use $RALength5_{it}$ (average length of the first five reviews) instead of $RALength10_{it}$ in this analysis. An F test rejected the hypothesis that the coefficients for *GroupA* (β = 0.036, $p < 0.05$) and *GroupB* (β = -0.041, $p = 0.056$) in column (5) were equal (F(1, 21955) = 8.21, $p < 0.01$). Thus, we also found support for H1b using the first 5 reviews on the page.

To compare data points in Groups A and B directly, we used the coarsened exact matching (CEM) algorithm in STATA (Blackwell et al., 2009; Iacus et al., 2012) to divide data points (app on a specific date) into stratas that were similar in rank in the current period and then incorporated strata level (instead of app level) fixed effects. Our purpose was to identify the differential impact of membership in Groups A and B on ranking in the next period for closely-ranked apps in the current period. We matched (coarsened) data points based on $ARank_{it}$ (100 equally spaced groups), $ARCount_{it}$ and $ADays_{it}$. In addition to $ARank_{it}$, we included $ARCount_{it}$ and $ADays_{it}$ in the matching process because the life-stage of the app (captured through age of the app) and the popularity of the app (captured through the number of ratings) can have significant impact on the apps ranking in the next period. The algorithm matched 1,440 data points (out of 3946 in Groups A and B) into 138 strata with 832 data points from Group A and 608 data points from group B; strata that did not have data points in both groups were dropped. Thus, data points in each strata were closely matched in current app ranking (the maximum difference in current app rank among data points in the same strata was only 15 while the maximum difference in rank in the sample was 1500), the age of the app and the number of ratings for the app. Maximum within-strata differences for other continuous variables were as follows (the corresponding maximum full-sample differences in parenthesis): $ARCount_{it}$ 3337 (104407); $APrice_{it}$ 20 (70); $ADispersion_{it}$ 1.2 (2); $ADays_{it}$ 150 (1874); $RALength10_{it}$ 121 (236). Thus, apps within each strata were well-matched with large reductions in within-strata variation for all variables compared to the corresponding within-sample variation. We then evaluated the following model where U_k is the strata level fixed effect and *GroupAvsB* is a dummy variable that is 1 if the data point belonged to Group A and 0 if the data point belonged to Group B; Group C was excluded from this analysis.

$$Ln(ARank_{it+1}) = \beta_0 + \beta_1 ARCount_{it} + \beta_2 APrice_{it} + \beta_3 AUpd_{it} + \beta_4 ADispersion_{it} + \beta_5 ADays_{it} + \beta_6 RALength10_{it} + \beta_7 GroupAvsB + U_k + \in_{it} \quad (3)$$

The results are shown in Table 6, columns (6) and (7). Column (6) includes only the control variables. Since a strata may contain data points from multiple apps, the coefficients for the control variables can capture differences between apps and were different from columns (1) - (4). Thus, older and more established apps had lower numerical ranks (more downloads). Interestingly, higher-priced apps also had a slightly lower numerical rank and more downloads.

Column (7) introduces the *GroupAvsB* variable that is of primary interest to us. The coefficient of the *GroupAvsB* variable was positive and significant ($\beta = 0.083$, $p < 0.01$), indicating that data points in Group A had a higher numerical rank (fewer downloads) in the next period when compared to data points in Group B (note that data points in the same strata were closely matched on current rank). Thus, our results supported H1b. Column (8) repeats the analysis in column (7) using the first 5 reviews on a review page. The coefficient of the re-coded *GroupAvsB* variable was significant and positive ($\beta = 0.046$, $p < 0.05$), indicating support for H1B when we used the first 5 reviews on the review page.

It is important to note that an app's rank is based primarily on downloads; for example, Garg and Telang (2013) find a strong relationship between app rank and downloads. Although Apple does not disclose the exact algorithm, the rank may also incorporate the product's summary rating statistics (such as its overall average rating and number of ratings). However, it is very unlikely that the algorithm considers the valence of reviews on the first review page (which is constantly changing) in determining the rank. Thus, our empirical setup may inflate the effect of summary rating statistics and is thus a conservative test of H1b.

Discussion

To examine the relative impacts of the average rating and top reviews on consumers' purchase decisions proposed in H1a and H1b, we conducted an empirical study using actual ratings and reviews collected from Apple's App Store. The results of this analysis provided evidence that in situations where there is a trade-off between the average rating of the product and the valence of the top reviews, consumers' purchase (download) decisions are more influenced by individual reviews than the average rating of the product (app) as we proposed in H1b.

This initial study had a number of limitations. First, while we controlled a number of variables that can affect product sales and included app-level and CEM strata-level fixed effects, unobserved consumer-level individual differences (such as their intentions to read online reviews) that correlate with both the independent and dependent variables may cause additional endogeneity concerns. Second, although it is reasonable to assume that most consumers would not change the default order of an app's displayed reviews, some consumers may have. Third, the use of archival data precluded us from exploring the possible sources of the swaying effect. We conducted two experiments to address these limitations. The

randomized experiments allow us to eliminate potential endogeneity issues more conclusively and better understand the process underlying the swaying effect of individual reviews.

STUDY 2

In the second study, we conducted an experiment to examine the competing hypotheses in a more controlled setting. Specifically, in a hypothetical online decision-making task, participants read two products' rating profiles (i.e., the average and the number of product ratings) and the 3 most recent reviews and then made purchase decisions between the two products. We manipulated both the average rating and the valence of individual reviews through the two products within-subjects to account for endogeneity concerns of consumer-level individual differences. We varied the average product rating at two levels between the two products: 4 stars vs. 4.5 stars. Meanwhile, we varied the valence of individual reviews (more positive vs. more negative) in such a way that one product would be superior based on the average rating, while the other product would be superior based on individual reviews. Specifically, the 4-star product had 2 positive and 1 negative reviews, and the 4.5-star product had 1 positive and 2 negative reviews (see Table 7). As a reminder, this unique within-subjects design is a common paradigm to explore the relative weights of different attributes in consumer decision-making (Ajzen, 2008). Consumers' preferences between the two alternatives would reflect the importance of the attributes (e.g., Shiv and Fedorikhin, 1999). In our case, a rational participant should choose the 4.5-star product over the 4-star product (as we proposed in H1a) because the average product rating is a more comprehensive signal of product quality based on hundreds of reviews. However, the ease-of-retrieval explanation dictates that participants should prefer the 4-star product with relatively more positive individual reviews (as we proposed in H1b) because reviews are more concrete and accessible.

Table 7: Two-Alternative Two-attribute Design in Study 2

	Attribute 1: Average Rating	**Attribute 2: Average Valence of Individual Reviews**
Alternative X	4-star	2 positive + 1 negative
Alternative Y	4.5-stsar	1 positive + 2 negative
	Y is better according to Attribute 1	X is better according to Attribute 2

Stimulus Materials

We used the digital camera as our context because it is a familiar product for most people. To vary the valence of 3 individual reviews of each product, we developed 6 sets of treatment reviews (with a positive version and a negative version in each set) to be used as a source of reviews for the two products. We started with three common camera features—ease of use, LCD, and image stabilization—and then wrote 4 sets of individual reviews for each feature (12 sets in total) based on the reviews used in Liu and Karahanna (2017) and actual camera reviews from Amazon. Within each review set, we first prepared a positive version and then constructed a corresponding negative version by adding negations and using

antonyms while holding the substantial content identical. Because we also kept the number of words in each review at around 25, the only difference between the positive and negative versions within each set is the valence.

To remove possible confounds, we conducted a pretest to ensure that two versions within each review set are equivalent in extremity and that different review sets are equivalent in terms of information quantity, concreteness, extremity, helpfulness, emotional intensity, realism, and reading difficulty. We recruited 55 participants from Amazon Mechanical Turk (MTurk) and asked them to read and evaluate 12 reviews, one review at a time. We randomly assigned them to read one version (either positive or negative) of the reviews in each review set. After reading each review, we asked each participant to report their evaluations of the review's 1) extremity (e.g., "not at all positive / very positive"), 2) information quantity (e.g., "contains very little information / contains a great deal of information"), 3) concreteness (e.g., "not at all concrete / very concrete"), 4) helpfulness (e.g., "not at all helpful / very helpful"), 5) emotional intensity (e.g., "contains little emotion / contains a great deal of emotion"), 6) realism (e.g., "not at all realistic / very realistic"), and 7) reading difficulty (e.g., "very hard to read / very easy to read"). Each variable was measured by two items adapted from prior literature along a 9-point scale (see Appendix B in E-Companion for all the measures and their sources).

We next conducted independent-samples t-tests of extremity between the positive and negative versions in each review set and paired-samples t-tests of all other variables (e.g., information quantity, concreteness, extremity, etc.) across different review sets. Based on the pretest results, we selected 6 sets of reviews (see Table 8) to be used for the two treatment products that satisfied the following criteria: each review set selected for a particular product described a different feature; the positive and negative versions within each set were equivalent in extremity (*t*s <= 1.589, *p*s >= 0.118); the positive and negative versions across different review sets were equivalent in all other relevant variables (*t*s <= 1.344 and 1.705, *p*s >= 0.191 and 0.100). Therefore, the 3 sets of treatment reviews we chose for each product are not significantly different in extremity between two versions within each review set or in other relevant variables (e.g., information quantity, concreteness, extremity, etc.) across different review sets.

Table 8: Three Sets of Reviews for Each Product

Set #	Review Source 1 (Used for One Product)	
	Positive Version	**Negative Version**
1	This camera is user-friendly compared to other entry level cameras. After just a few days of use, I found it really intuitive.	This camera is not user-friendly compared to other entry level cameras. After just a few days of use, I found it really complicated.
2	The camera has an excellent LCD screen, which is large enough for most people. It works well in both high and low light conditions.	The camera has a poor LCD screen, which is not large enough for most people. It doesn't work well in either high or low light condition.

3	The image stabilization works as expected. It can correct the impact of minor accidental hand motion. The pictures taken in unsteady situations are clear.	The image stabilization doesn't work as expected. It fails to correct the impact of accidental hands motion. The pictures taken in unsteady situations are blurry.
	Review Source 2 (Used for the Other Product)	
	Positive Version	**Negative Version**
1	Everything on this camera is easy to use. It's a good choice for people who don't have much experience with digital cameras.	Everything on this camera is hard to use. It's a bad choice for people who don't have much experience with digital cameras.
2	The LCD screen has quick feedback. After I take shots, I can see the pictures flashed onto the LCD screen instantly.	The LCD screen has slow feedback. After I take shots, I can see the pictures flashed onto the LCD screen after a while.
3	The image stabilization is effective. This feature ensures the clarity of photo if the camera is slightly moved when I take the photo.	The image stabilization is not effective. This feature cannot ensure the clarity of photo if the camera is slightly moved when I take the photo.

Procedure

Fifty-three undergraduate students (20 male) from a U.S. university took part in the experiment in exchange for extra credit.[5] Among the participants, 91 percent were originally from the U.S., 70 percent were juniors or above, and the average age of the students was 20.

In the cover story, participants were asked to imagine that they were planning to purchase a digital camera from Amazon.com, and their search returned two different digital cameras with the same price of $549.99. Then they were asked to read the rating profile (including the average product rating and the number of ratings from prior users) and the 3 most recent reviews displayed on the first review page for both cameras, one product at a time.[6]

Both cameras had accumulated hundreds of reviews, but they had different average product ratings (4 and 4.5 out of 5 stars). We counterbalanced the order of the treatment (i.e., which product has a higher average rating) and the order of the two review sources (see Table 8) assigned to the two products (with the left one labeled as "Product A" and the right one labeled as "Product B"). We also randomized the order of three reviews for each product. In addition, we varied the average valence of the 3 most recent reviews in such a way that the product superior in the average product rating is inferior in the average valence of individual reviews (see Table 7). In addition, for each product, we selected 3 reviews from the

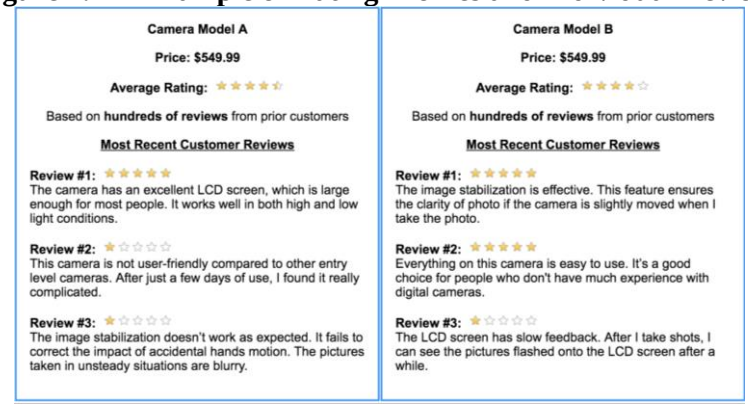
[5] Since both average product rating and valence of individual reviews were manipulated within-subjects, a sample size of 35 (or more) is sufficient to capture a repeated-measure effect of at least moderate size (f = .25) with 80% power (Faul et al., 2007).

[6] We used the term "most recent reviews" because consumers are less likely to read earlier/older reviews in reality, and they are more likely to read the most recent ones. However, the most observable reviews can be the most recent reviews, the most helpful reviews, or any other types of reviews depending on the default sorting order of reviews on different platforms. To examine whether our findings hold with different sorting methods, we conducted a supplementary experiment (N = 167 undergraduate students) very similar to Study 2, except that the reviews were introduced as either the most helpful reviews or the randomly selected reviews. The results were consistent with those of Study 2, suggesting that the swaying effect of individual reviews holds regardless of the sorting method.

3 review sets (of either review source 1 or 2 in Table 8), one version from each set, so that the positive and negative versions within the same set would not be presented to the same participant. Each of the 3 review sets also had an equal chance of being selected to represent a positive or negative review. We also displayed the corresponding rating (5-star for positive review and 1-star for negative review) along with the textual content of reviews.[7]

After observing the rating profile and reading the 3 most recent reviews of each product, participants were asked to report their intention to purchase the product using a 9-point scale adapted from Dodds et al. (1991) and Goldberg and Gorn (1987) (e.g., "If you were thinking of buying a digital camera, how likely is it that you would buy Camera Model A?"). Then participants were presented with the rating profiles and individual reviews of two products side by side (see Figure 2 for an example). The first product they evaluated earlier appeared on the left side of the screen. The product on the left could have an overall average rating of either 4 or 4.5 stars (due to the counterbalancing of the two products' order), thus mitigating a potential confound of the location effect. After observing the rating profiles and individual reviews of two products, participants were asked to choose one camera between the two options for purchase, using an 8-point scale (1 = "definitely choose Camera A", 8 = "definitely choose Camera B"). We used an 8-point scale here so that participants cannot keep a neutral stance between the two options. As a manipulation check, participants were also asked to recall the average rating of each product and to evaluate the valence of the 3 individual reviews as a whole for each product using three items adapted from MacKenzie and Lutz (1989) (e.g., "expresses very bad feelings about the camera / expresses very good feelings about the camera"; see Appendix C in E-Companion for all measures used in this study).

Figure 2: An Example of Rating Profiles and Individual Reviews



[7] We used 1-star and 5-star reviews because a) it is hard to construct a less extreme review with both positive and negative statements while also keeping the review rating at a consistent, less extreme level (e.g., 3 or 4 stars) and b) it is not uncommon in reality that the first few reviews displayed on the review platforms can be very extreme, either very positive or very negative.

Results

We first conducted manipulation checks for the two variables we manipulated in the study. The mean of participants' recalled average rating of the 4-star product was significantly lower than that of the 4.5-star product (M = 3.84 vs. 4.06, $F(1, 52) = 5.183$, $p = 0.028$), indicating that our manipulation of the products' overall average rating was successful and in the expected direction. Also, the perceived valence of the most recent reviews of the 4-star product (2 positive and 1 negative reviews) was significantly more positive than that of the 4.5-star product (1 positive and 2 negative reviews) (M = 6.63 vs. 3.50, $F(1, 52) = 115.489$, $p < 0.001$), indicating that our manipulation of the average valence of the most recent reviews was also successful.

To explore the relative impacts of the average product rating and the most recent reviews on consumers' purchase intention, we conducted a repeated-measure ANCOVA with the two products entered as a within-subjects factor. We added the order of treatment and the order of review set assignment as two covariates. Results revealed that participants' intention to purchase the 4-star product was significantly higher than their intention to purchase the 4.5-star product (M = 6.15 vs. 3.58, $F(1, 50) = 73.559$, $p < 0.001$). Although the 4.5-star product is superior to the 4-star product based on the average rating, participants' purchase intention of the 4-star product (whose individual reviews are more positive) was higher than the 4.5-star product, indicating that individual reviews swayed their purchase intentions from the 4.5-star to 4-star product. Thus, we obtained consistent evidence for the swaying effect of individual reviews (H1b) as in Study 1.

In addition, we investigated whether the swaying effect of individual reviews also shaped consumer choice. Participants provided their choice between the two product options along an 8-point scale (1 = "definitely choose Camera Model A," 8 = "definitely choose Camera Model B"). We re-coded the choice values so that a lower value indicates participants' preference for the 4.5-star product and a higher value indicates their preference for the 4-star product. Then we conducted a one-sample t-test to compare the mean of participants' choices with the midpoint (4.5) of the scale. Results revealed that the mean value of re-coded responses was 6.32, which was significantly above the midpoint ($t(52) = 8.970$, $p < 0.001$).[8] Thus, participants preferred the 4-star product (with more favorable reviews) to the 4.5-star product in their choice, providing additional evidence for a swaying effect of individual reviews.

Discussion

Study 2 examined the relative impacts of the average product rating and individual reviews on consumers' purchase decisions through a carefully designed experiment. Consistent with the findings of

[8] As a robustness check, we conducted an ANCOVA analysis with consumers' choice versus the midpoint (4.5) entered as repeated measures of the outcome variable, and we controlled for treatment order and review set order as covariates. We found consistent results with the one-sample t-test reported in the main text.

Study 1, the results indicated that a few individual reviews could sway consumers' purchase decisions, providing additional evidence for H1b.

Our design in Study 2 still had several limitations. First, although the results of manipulation checks indicated a successful manipulation of the average ratings and the average valence of individual reviews, the swaying effect of individual reviews might result from the greater prominence of individual reviews on the page compared to the average ratings. Specifically, the average ratings of the two products utilized in this study were very positive (i.e., 4 and 4.5 stars), with a small difference of 0.5 stars. While a 0.5-star difference in the average product rating based on hundreds of reviews is substantial, this 0.5-star difference at the positive end of the rating scale might be less noticeable in the eyes of consumers. Second, we did not reveal the exact total number of prior reviews for each product and simply displayed "hundreds of reviews." While this wording for review volume was held identical, there was some ambiguity regarding the representativeness of the 3 most recent individual reviews between the two products (e.g., 3 out of 100 vs. 3 out of 900 prior reviews). Although we deem this differential interpretation unlikely, revealing the exact numbers for review volume could eliminate this possibility. Third, consumers' purchase decisions in this study might have been swayed by a recency effect: participants saw the individual reviews in the end, immediately before they answered the purchase intention question. Finally, we focused on comparing the main effects of the average product rating and individual reviews, but we did not explore the possible source of the swaying effect of individual reviews. Because we displayed the textual content and rating score of individual reviews simultaneously, we could not determine whether the swaying effect of individual reviews was driven by the concrete review text or the review rating. We designed the final experiment to address these limitations.

STUDY 3

In Study 3, we utilized a similar design as Study 2 with a few exceptions. First, to increase the salience of average product ratings, we manipulated the average ratings of the two products to have a greater difference (2.5 vs. 3.5 stars) and increased the size and prominence of their rating stars and associated wordings.[9] Second, we displayed the exact total number of reviews (125 vs. 127; counterbalanced in order and not significantly different from each other) to ensure the equivalence of the representativeness of the 3 most recent reviews. Third, we explored the source of the effect of individual reviews by using a "moderation-of-process" design, in which we manipulated the potential process variable directly (Spencer et al., 2005).

[9] We manipulated the average ratings to be less positive than those used in Study 2 because of two reasons. First, the lowered average ratings are more comparable to the difference between the average valence of individual reviews (2.3 vs. 3.7 stars). Second, such a scenario could occur in reality when consumers really need a certain type of product but very few options are available on the market.

Following our earlier arguments based on the mere-accessibility framework and ease-of-retrieval explanation, consumers should rely more on accessible and easy-to-recall information in their decision-making and choices. While several factors can contribute to the accessibility of information, the extent to which the information is concrete has been considered one of the most important factors (Nisbett and Ross, 1980). The reason is that concrete information can increase the number of associative routes in memory that imply a specific concept (termed as cognitive elaboration; see Anderson and Bower, 1980; Nisbett and Ross, 1980) and thus increase the accessibility of information (Kisielius and Sternthal, 1986). When processing more concrete information, a greater number of associative routes would be established and evoked in human memory through which relevant information could be retrieved and recalled (Kisielius and Sternthal, 1986). In addition, experimental studies examining concreteness have manipulated this factor through narrative information versus statistical information or numerical ratings (e.g., Keller and Block, 1997). Because narrative information is easier to stimulate the cognitive elaboration of relevant associations in memory and subsequently easier to imagine and recall than statistical information or numerical ratings, the former comes to mind more easily and affects decisions to a greater extent than the latter. Applied to our context, since the textual content of individual reviews is more concrete than their associated ratings, the swaying effect of individual reviews could be driven by the textual content rather than the ratings of reviews.

In this study, we adopted a "moderation-of-process" design because the possible source of the effect of individual reviews—concreteness of reviews' textual content versus their ratings—can be easily manipulated but hard to measure directly. In this "moderation-of-process" design, researchers can explore a potential mechanism by manipulating it directly as an additional variable (besides the independent variable) in the experiment and then testing the moderating effect of this additional variable (Spencer et al., 2005). A significant moderation provides evidence for the mechanism because the effect of the independent variable could be weakened or "turned off" when the mechanism is not in place. In our setting, we varied rating vs. text between-subjects, presenting only the ratings of individual reviews to half of the participants and only the concrete content of individual reviews to the other half. If the observed effect of individual reviews (relative to average product ratings) is significantly different between the two conditions (i.e., the effect is moderated by concreteness), this result would suggest concreteness as a possible source of the swaying effect of individual reviews and also rule out recency effect as an alternative explanation. Individual reviews are presented below the average product ratings regardless of the condition (i.e., rating vs. text), so a recency effect would predict a lack of moderation. Taken together, we propose an additional hypothesis below.

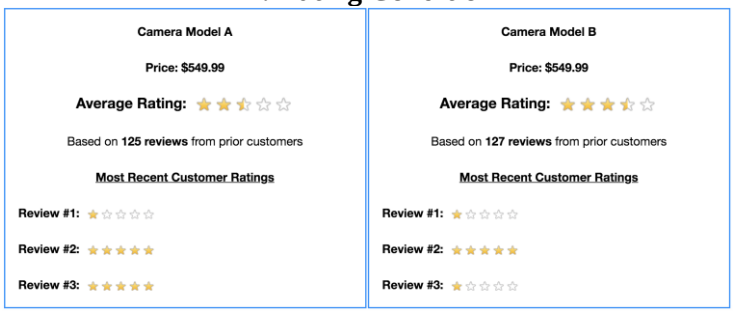
Hypothesis 2: Given a choice set of product options, the influence of individual reviews (vs. average ratings) on consumers' purchase decisions is mediated by the reviews' textual content.

Procedure and Measures

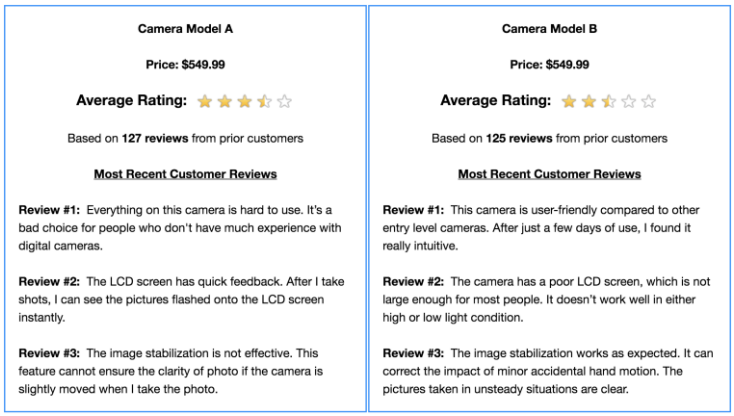
Seventy-three undergraduate students (41 male) from a U.S. university took part in this experiment in exchange for extra credit. Among them, 70 percent were originally from the U.S., 70 percent were juniors or above, and the average age of the students was 22. This study followed a procedure similar to Study 2, except that each subject was randomly assigned to the rating-only or text-only condition. Those in the rating-only condition only saw the rating (1 or 5 stars) of individual reviews, while those in the text-only condition only saw the textual content (see Figure 3 for an example). As a manipulation check, we asked participants to report the concreteness of the review information for each product with six items adapted from Keller and Block (1997) (e.g., "not at all concrete / very concrete") and to evaluate the average valence of the 3 individual reviews as a whole for each product using the same items as Study 2. See Appendix C in E-Companion for the measures used in this study.

Figure 3: An Example of Rating Profiles and Individual Reviews

A. Rating Condition



B. Text Condition



Results

First, we conducted manipulation checks for the average product rating, perceived valence of the most recent reviews, and perceived concreteness. Results revealed that the recalled average rating of the 2.5-star product was significantly lower than that of the 3.5-star product (M = 2.76 vs. 3.22, $F(1, 71)$ = 23.380, $p < 0.001$), and that perceived average valance of 2 positive and 1 negative reviews was significantly higher than that of 1 positive and 2 negative reviews (M = 5.97 vs. 3.60, $F(1, 71)$ = 48.947, p

< 0.001). Also, the perceived concreteness in the rating-only condition was significantly lower than that in the text-only condition (M = 3.83 vs. 6.17, $F(1, 71) = 48.949$, $p < 0.001$). Therefore, the manipulations of all the variables were deemed successful.

Next, to explore the swaying effect of individual reviews on consumers' purchase decisions (H1b) and its source (H2), we conducted ANCOVA with participants' intention to purchase the product entered as the dependent variable, the two products entered as a within-subjects factor, and concreteness (rating-only vs. text-only) entered as a between-subjects factor. We also added treatment order, review set order, and review volume order as covariates. Results revealed that participants' intention to purchase the 2.5-star product was significantly higher than their intention to purchase the 3.5-star product (M = 4.88 vs. 3.72, $F(1, 68) = 10.529$, $p = 0.002$), providing additional evidence for H1b. In addition, the interaction between the effect of individual reviews and concreteness was marginally significant ($F(1, 68) = 3.725$, $p = 0.058$). Pairwise comparisons revealed that when participants were presented with the textual content of individual reviews, their intention to purchase the 2.5-star product was significantly higher than their intention to purchase the 3.5-star product (M = 5.44 vs. 3.58, $F(1, 68) = 12.846$, $p = 0.001$). However, when participants were presented with only the ratings of individual reviews, the difference in their purchase intentions between the two products was not significant ($F(1, 68) = 0.906$, $p = 0.345$). Together, these results indicated that the swaying effect of individual reviews was driven by the concrete textual content of reviews, not by their ratings or a recency effect.

Moreover, we investigated the source of the swaying effect of individual reviews on participants' choice between the two product options. Following a similar analysis in Study 2, we re-coded participants' choice, with a value above the midpoint (4.5) indicating a preference for the 2.5-star product and a value below the midpoint indicating a preference for the 3.5-star product. We conducted a one-sample t-test for rating-only and text-only conditions, respectively. When the textual content of individual reviews was displayed, the mean value of consumers' choice (M = 5.34) was significantly above the midpoint ($t(34) = 2.257$, $p = 0.031$), indicating that participants preferred the 2.5-star product to the 3.5-star product in this condition. However, when the ratings of individual reviews were displayed, the mean value of consumers' choice (M = 4.03) was not significantly different from the midpoint ($t(37) = 1.087$, $p = 0.284$).[10] Hence, these results offered additional evidence that the concrete review texts might be the source of the swaying effect of individual reviews.

[10] As a robustness check, we conducted ANCOVA with consumers' choice versus 4.5 (the midpoint of the scale) entered as repeated measures of the outcome variable and concreteness (rating-only vs. text-only) entered as a between-subjects factor. We controlled for treatment order, review set order, and review volume order as covariates. Results revealed that the interaction between the within-subjects and between-subjects factors was significant ($F(1, 68) = 5.077$, $p = 0.027$). Pairwise comparisons of the mean value of consumers' choice versus the midpoint under rating-only and text-only conditions were consistent with the results of one-sample t-tests.

Discussion

In Study 3, we examined the source of the swaying effect of individual reviews. The study replicated the swaying effect of individual reviews (H1b) when average product ratings were closer to neutral, more distant from each other, and more prominently displayed. It ruled out the recency effect as an alternative explanation. More importantly, our findings suggested the concrete textual content rather than the ratings of individual reviews as a possible source of the swaying effect of individual reviews (H2).

CONCLUSIONS

The online word-of-mouth literature usually assumes that a product's average rating and other summary rating statistics are the primary drivers of purchase decisions and product sales (see Babić Rosario et al., 2016; Floyd et al., 2014). However, emerging evidence suggests that individual reviews also play an important role in consumer purchase decisions (e.g., Liu et al., 2019; Vana and Lambrecht, 2021; Yin et al., 2021). We focus on consumers who consult both average ratings and individual reviews and propose two competing hypotheses regarding their *relative* importance in purchase decisions. To disentangle their effects, we adopted a trade-off design and examined consumers' purchase preferences between product options whose average ratings contradict individual reviews. Through one archival and one experimental study, we obtained evidence of a swaying effect of individual reviews. Additionally, the results of a follow-up experiment demonstrated the textual content of individual reviews as a possible source of the swaying effect.

Practical Implications

Our findings provide practical implications for product manufacturers, retailers, and review platforms. First, to keep track of online reputation and consumer comments, product manufacturers and retailers often ask consumers to leave reviews on either business or third-party review sites. A key implication for product manufacturers and retailers is that their prevailing focus on average ratings and other summary rating statistics (vs. individual reviews) may not be the most optimal strategy. Although it is widely accepted that average ratings and other summary rating statistics play a greater role than individual reviews in consumers' purchase decisions and product sales, the swaying effect we demonstrate implies that consumers who consult both average ratings and individual reviews tend to place more emphasis on a few top-ranked reviews than on the average product rating. Therefore, consumers' pre-decision impression of products may be closer to the consumption experiences shared in top-ranked reviews. Thus, product manufacturers and online retailers may be misguided if they gauge consumer interest and purchase intentions based primarily on average ratings and other summary rating statistics. Further, while the average product rating is not easy to change because it is calculated based on all historical ratings, the first few reviews that consumers observe can change over time. Our results suggest

that product manufacturers and online retailers can better influence consumers' purchase decisions (and hence stimulate sales) by being more strategic in dealing with these two kinds of information. For example, consumers may be turned away from a product if it has a low average rating, but a higher average rating is not sufficient to result in a purchase if the top-ranked reviews that consumers see are not positive.

In addition, our findings provide important practical insights for manufacturers and retailers who intend to enhance their marketing strategies based on online word-of-mouth. Because high average ratings are no guarantee for "success," businesses should be keenly aware of those highly accessible reviews that often deviate from the average ratings and can sway consumers' ultimate purchase decisions. Moreover, businesses can incorporate the swaying effect of top reviews into their marketing strategies by focusing on the most helpful or most recent reviews, whichever sorting method is applied by default and thus more likely to influence consumers' decisions. For instance, businesses aiming at addressing consumers' concerns and negative comments can prioritize responding to top-ranked negative reviews. While the average ratings are harder to change and also less influential for serious consumers, businesses' effective responses to top-ranked negative reviews can instantly attenuate future consumers' negative inferences and reduce the likely swaying effect of those highly accessible negative opinions (Gu and Ye, 2014).

Second, our investigation into the possible source of the swaying effect suggests that the influence of individual reviews on consumers' purchase decisions could be driven more by the concrete textual content of reviews than the ratings of reviews. Thus, review platforms should provide greater incentives for consumers to write text reviews that describe their experience with the product rather than simply providing a rating. As an example, Steam, a digital game delivery platform, displays only the textual content (without the rating) of individual reviews for potential users to read because concrete opinions expressed in the reviews are what users pay attention to. In addition, to reveal "the wisdom of the crowd" and prevent over-reliance on idiosyncratic reviews, review platforms may consider displaying the most helpful reviews or most recent reviews in a way that is aligned with the product's average rating.

Third, when review platforms design the layout of product pages, they could benefit consumers by spotlighting individual reviews, and the current strategy of universally displaying the average ratings in the most prominent places may not be effective. Consumers who consult both average ratings and individual reviews rely more on the latter in their purchase decisions. Thus, our research suggests that review platforms should incorporate individual reviews into the design of product pages. For example, review platforms may consider displaying a few top-ranked reviews in more salient places on the product pages in addition to average ratings. Moreover, aggregating information from the top-ranked reviews (e.g., the most helpful reviews, the most recent reviews) might be another way for review platforms to facilitate consumers' decision-making. For example, along with listing the most recent or most helpful

reviews, review platforms can display frequently mentioned keywords from top reviews in salient places. Review platforms should also design better sorting methods for the reviews and allow consumers to easily adjust the sorting order of the presented reviews based on their personal preferences. They should also be more thoughtful in determining the reviews' default order, as that is the order used by most consumers.

Limitations and Future Research

This study also has several limitations that provide opportunities for future research. First, we focused on the average rating in this paper, which is the most salient signal of product quality before consumers read individual reviews. However, other information cues may also affect consumers' impressions about a product and their purchase decisions, such as the product's brand image and the description of product features on the website. Future research can explore the role of these alternative information cues.

Second, although we demonstrate that the swaying effect may be driven by the detailed review content (vs. review ratings), the underlying mechanisms of the swaying effect are not fully explored in this research. For example, to develop a more comprehensive understanding of the swaying effect, it might be valuable to investigate consumers' motivation to search for and read individual reviews. In addition, consumers' self-determined number of reviews they choose to read may also play a role in the swaying effect. Future research can also explore other boundary conditions (e.g., product features and review characteristics) that can enhance or weaken the swaying effect. Further, the average rating and other summary rating statistics should play a more important role in initial product search and discovery. Future research can explore the impact of summary rating statistics on review seeking and consumption at earlier stages of consumer decision-making.

Third, in our created scenarios in the experimental studies, the displayed reviews were labeled as "most recent reviews" on the first review page, and they were the only reviews that consumers read during the experiments. We did not offer participants the ability to change the order of reviews to achieve tight control and avoid compromising the internal validity of the experiments. On the other hand, real-world consumers can change the display order of reviews. As a result, the first few reviews that are the most accessible to consumers can also be the most recent, the most helpful, the most favorable, or the most critical reviews. Although this concern does not apply to our experimental studies, future research can study whether the swaying effect demonstrated in this paper depends on the display order of reviews.

ACKNOWLEDGMENTS

We appreciate the department editor, the senior editor, and three anonymous reviewers for their constructive guidance, comments, and suggestions throughout the review process. We thank Qianqian Ben Liu for sharing the review stimuli from his (2017) paper co-authored with Elena Karahanna, which we used to construct our review stimuli in the experiments. We are also grateful to Marius Florin

Niculescu, Lizhen Xu, Michael Smith, and Adrian Gardiner for their help in recruiting experiment participants.

REFERENCES

Ajzen, I. 2008. Consumer attitudes and behavior. In: Haugtvedt, C.P., Herr, P.M., Kardes, F.R.K. (Eds), *Handbook of Consumer Psychology*. Routledge, New York, 525-548.

Anderson, J.R., Bower, G.H. 1980. *Human associative memory*. Lawrence Erlbaum, Hillsdale, NJ.

Ba, S., Jin, Y., Li, X., Lu, X. 2020. One size fits all? The differential impact of online reviews and coupons. *Production and Operations Management*. 29(10): 2403-2424.

Babić Rosario, A., Sotgiu, F., De Valck, K., Bijmolt, T.H.A. 2016. The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*. 53(3): 297-318.

Banerjee, A.V. 1992. A simple model of herd behavior. *The Quarterly Journal of Economics*. 107(3): 797-817.

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.

Blackwell, M., Iacus, S.M., King, G., Porro, G. 2009. CEM: Coarsened exact matching in Stata. *Stata Journal*. 9(4): 524-546.

Cai, H., Chen, Y., Fang, H. 2009. Observational learning: Evidence from a randomized natural field experiment. *American Economic Review*. 99(3): 864-882.

Ceran, Y., Singh, H., Mookerjee, V. 2016. Knowing what your customer wants: Improving inventory allocation decisions in online movie rental systems. *Production and Operations Management*. 25(10): 1673-1688.

Chen, H., Zheng, Z., Ceran, Y. 2016. De-biasing the reporting bias in social media analytics. *Production and Operations Management*. 25(5): 849-865.

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. 2015. Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*. 42(6): 817-833.

Dodds, W.B., Monroe, K.B., Grewal, D. 1991. Effects of price, brand, and store information on buyers' product evaluations. *Journal of Marketing Research*. 28(3): 307-319.

Duan, W., Gu, B., Whinston, A.B. 2008a. Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems*. 45(4): 1007-1016.

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

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*. 39(2): 175-191.

Floyd, K., Freling, R., Alhoqail, S., Cho, H.Y., Freling, T. 2014. How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing*. 90(2): 217-232.

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

Garg, R., Telang, R. 2013. Inferring app demand from publicly available data. *MIS Quarterly*. 37(4): 1253-1264.

Goldberg, M.E., Gorn, G.J. 1987. Happy and sad TV programs: How they affect reactions to commercials. *Journal of Consumer Research*. 14(3): 387-403.

Gu, B., Ye, Q. 2014. First step in social media: Measuring the influence of online management responses on customer satisfaction. *Production and Operations Management*. 23(4): 570-582.

Guan, X., Wang, Y., Yi, Z., Chen, Y.J. 2020. Inducing consumer online reviews via disclosure. *Production and Operations Management*. 29(8): 1956-1971.

Gutt, D., Neumann, J., Zimmermann, S., Kundisch, D., Chen, J. 2019. Design of review systems–A strategic instrument to shape online reviewing behavior and economic outcomes. *The Journal of Strategic Information Systems*. 28(2): 104-117.

Higgins, E.T. 1996. Knowledge activation: Accessibility, applicability, and salience. *Social Psychology: Handbook of Basic Principles*, 133-168.

Hu, N., Zhang, J., Pavlou, P.A. 2009. Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*. 52(10): 144-147.

Iacus, S.M., King, G., Porro, G. 2012. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*. 20(1): 1-24.

Jabr, W., Liu, B., Yin, D., Zhang, H. 2020. Online word-of-mouth. In: Bush, A., Rai, A. (Eds), *MIS Quarterly Research Curations*.

Jabr, W., Rahman, M.S. forthcoming. Online reviews and information overload: The role of selective, parsimonious, and concordant top reviews. *MIS Quarterly*.

Keller, P.A., Block, L.G. 1997. Vividness effects: A resource-matching perspective. *Journal of Consumer Research*. 24(3): 295-304.

Kisielius, J., Sternthal, B. 1986. Examining the vividness controversy: An availability-valence interpretation. *Journal of Consumer Research*. 12(4): 418-431.

Lau, R.Y.K., Zhang, W., Xu, W. 2018. Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Production and Operations Management*. 27(10): 1775-1794.

Lei, Z., Yin, D., Zhang, H. 2021. Focus within or on others: The impact of reviewers' attentional focus on review helpfulness. *Information Systems Research*. 32(3): 801-819.

Liu, Q.B., Karahanna, E. 2017. The dark side of reviews: The swaying effects of online product reviews on attribute preference construction. *MIS Quarterly*. 41(2): 427-448.

Liu, X., Lee, D., Srinivasan, K. 2019. Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *Journal of Marketing Research*. 56(6): 918-943.

MacKenzie, S.B., Lutz, R.J. 1989. An empirical examination of the structural antecedents of attitude toward the ad in an advertising pretesting context. *The Journal of Marketing*. 53(2): 48-65.

Mafael, A., Gottschalk, S.A., Kreis, H. 2016. Examining biased assimilation of brand-related online reviews. *Journal of Interactive Marketing*. 36: 91-106.

Menon, G., Raghubir, P. 2003. Ease-of-retrieval as an automatic input in judgments: A mere-accessibility framework? *Journal of Consumer Research*. 30(2): 230-243.

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

Nisbett, R.E., Ross, L. 1980. *Human inference: Strategies and shortcomings of social judgment*. Prentice-Hall, Englewood Cliffs, NJ.

PowerReviews. 2021. Survey: The Ever-Growing Power of Reviews. Retrieved from https://www.powerreviews.com/insights/power-of-reviews-survey-2021/ (accessed date: December 13, 2021).

Qahri-Saremi, H., Montazemi, A.R. 2019. Factors affecting the adoption of an electronic word of mouth message: A meta-analysis. *Journal of Management Information Systems*. 36(3): 969-1001.

Roberts, J.H., Lattin, J.M. 1991. Development and testing of a model of consideration set composition. *Journal of Marketing Research*. 28(4): 429-440.

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.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., Simons, A. 1991. Ease of retrieval as information: another look at the availability heuristic. *Journal of Personality and Social Psychology*. 61(2): 195.

Shen, W., Hu, Y.J., Ulmer, J.R. 2015. Competing for attention: An empirical study of online reviewers' strategic behavior. *MIS Quarterly*. 39(3): 683-696.

Shiv, B., Fedorikhin, A. 1999. Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research*. 26(3): 278-292.

Spencer, S.J., Zanna, M.P., Fong, G.T. 2005. Establishing a causal chain: why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*. 89(6): 845.

Sun, H., Xu, L. 2018. Online reviews and collaborative service provision: A signal-jamming model. *Production and Operations Management*. 27(11): 1960-1977.

Sun, M. 2012. How does the variance of product ratings matter? *Management Science*. 58(4): 696-707.

Taylor, S.E., Thompson, S.C. 1982. Stalking the elusive "vividness" effect. *Psychological Review*. 89(2): 155-181.

Tversky, A., Kahneman, D. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*. 5(2): 207-232.

Vana, P., Lambrecht, A. 2021. The effect of individual online reviews on purchase likelihood. *Marketing Science*. 40(4): 593-812.

Wang, Y., Goes, P., Wei, Z., Zeng, D. 2019. Production of online word-of-mouth: Peer effects and the moderation of user characteristics. *Production and Operations Management*. 28(7): 1621-1640.

Watson, J., Ghosh, A.P., Trusov, M. 2018. Swayed by the numbers: The consequences of displaying product review attributes. *Journal of Marketing*. 82(6): 109-131.

Wright, P., Barbour, F. 1977. *Phased decision strategies: Sequels to an initial screening*. Graduate School of Business, Stanford University.

Yan, L., Yan, X., Tan, Y., Sun, S.X. 2019. Shared minds: How patients use collaborative information sharing via social media platforms. *Production and Operations Management*. 28(1): 9-26.

Yin, D., Bond, S.D., Zhang, H. 2017. Keep your cool or let it out: Nonlinear effects of expressed arousal on perceptions of consumer reviews. *Journal of Marketing Research*. 54(3): 447-463.

Yin, D., Bond, S.D., Zhang, H. 2021. Anger in consumer reviews: Unhelpful but persuasive? *MIS Quarterly*. 45(3): 1059-1086.

Yin, D., Mitra, S., Zhang, H. 2016. Research note—When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth. *Information Systems Research*. 27(1): 131-144.

E-COMPANION

Appendix A: Literature Review on Online Reviews in Operations Management Research

Research	Study/Theory	Methods/Constructs	Main Findings
Gu and Ye (2014)	**Study:** Examined the impact of management responses to reviews on customer satisfaction not only for customers who receive the responses but also for those who observe the responses **Context:** Hotel reviews retrieved from Ctrip.com **Theory:** Social exchange theory and peer-induced fairness theory	**Empirical model:** Developed a panel data regression model that controls for regression towards the mean and heterogeneity in individual preferences **DV:** Hotel rating **IVs:** Whether a reviewer is aware of the management response, whether a reviewer observes the management response **Moderator:** Level of customer (reviewer) satisfaction	Online management responses are highly effective on customers who are unsatisfied with the service provider but have limited influence on other customers. Online management responses increase future satisfaction of the complaining customers who receive the responses and reduce future satisfaction for those who observe the responses to others but do not receive the responses.
Ceran et al. (2016)	**Study:** Investigated the circular relationship among three outcome variables: rental generates WOM, WOM creates desire to rent, and desire to rent turns into rental **Context:** Movie data collected from a movie rental company and the Internet Movie Database	**Empirical model:** - Developed a three-equation system with desire (number of subscribers who wish to rent), WOM (number of reviews), and rental (number of rentals) as the dependent variables in the corresponding equation - Used 3SLS to estimate the three-equation system	Rental has a positive effect on WOM, WOM can positively impact desire, and desire has a positive influence on rental.
Chen et al. (2016)	**Study:** Modeled the data generating process of UGC and rectified the reporting biases due to silent users who do not write a review **Context:** Movie reviews of Blockbuster.com	**Analytical model:** - Developed an integrated stochastic model to capture the underlying data generating process of UGC - Extended the Beta Binomial/ Negative Binomial Distribution framework to model a user's reporting process - Developed an inverse probability weighting method to rectify the reporting biases	The distribution of the reporting probability of a user who has a positive opinion towards a product stochastically dominates the one who has a negative opinion. Found underestimated influence of review volume and sentiment on sales when users' reporting biases are unaccounted for.
Lau et al. (2018)	**Study:** Designed big data analytic methods based on product reviews to improve sales forecasting	**Method:** Used a parallel aspect-oriented sentiment analysis algorithm to mine sentiments from product reviews to enhance sales forecasting performance	Their large-scale empirical test confirms that consumer sentiments mined from product reviews can improve sales forecasting.

Sun and Xu (2018)	**Study:** Explored the role of online reviews in the provision of collaborative services	**Analytical model:** Developed a signal-jamming model of the review generating and information updating processes	The client review (compared with service outcome) observed as a signal of provider type leads to lower effort by both the client and the provider. When private information about the provider type is incorporated in client reviews, service providers are better motivated to work diligently. When both the client review and the service outcome are available, service providers lack sufficient incentive to devote effort.
Wang et al. (2019)	**Study:** Investigated how the population size of audience influences the volume of reviews, the variance of ratings, and the helpfulness of reviews **Context:** Product reviews collected from Douban.com **Theory:** Theories of prosocial behavior and social pressure	**Empirical model:** - Conducted a quasi-experiment based on a difference-in-difference framework - Used the introduction of "Douban Reading" as an exogeneous shock **DVs:** Volume of reviews, variance of ratings, and review helpfulness	Found that larger audience causes individuals to write more reviews with higher quality and assign higher but also more diverse ratings.
Yan et al. (2019)	**Study:** Assessed the influence of shared health information and experiences on patients' perceived treatment outcome **Context:** Treatment reviews retrieved from an online healthcare community **Theory:** Social information processing theory	**Empirical model:** - Used a latent response model to explore the effect of shared information on patients' treatment ratings - Developed a structural model to examine the influence of shared information on perceived treatment outcome **DV:** Patients' treatment rating (perceived treatment outcome) **IVs:** Aggregated positive valence and rating dispersion of generalized others' reviews, aggregated positive valence and rating dispersion of familiar others' reviews	The rating dispersion of generalized others' treatment has a positive effect on patients' perceived treatment outcome, while the rating dispersion of familiar others' treatment has a negative influence. The former effect is stronger than the latter effect. The positive sentiment in others' treatment reviews negatively affects patients' perceived treatment outcome.

Ba et al. (2020)	**Study:** Examined the moderating effects of consumers' information search characteristics and product signaling characteristics on the role of online reviews in consumers' decision-making process **Context:** Restaurant reviews from a Chinese restaurant review site **Theory:** Information search theory and signaling theory	**Empirical model:** Conditional logit model **Moderators:** - Consumers' information search characteristics: inertia tendency, shopping experience, spending level, and coupon proneness - Product signaling characteristics: price level, coupon presence, and coupon frequency **DV:** Dining probability (product sales) **IVs:** Average rating of reviews (review valence) and number of reviews (review volume)	Review valence and review volume have distinct impacts on consumers with different information search characteristics based on inertia tendency, shopping experience, and coupon proneness, and they influence consumers' purchase decisions differently depending on product price level and product coupon promotions.
Guan et al. (2020)	**Study:** Investigated how consumers' reference-dependent preferences (with regard to product quality) and consumers' type (naïve vs. sophisticated) influence a seller's voluntary disclosure strategy to manipulate consumer reviews	**Analytical model:** Developed a dynamic model to explore the joint effect of consumers' type and their reference-dependent preferences	The seller can strategically manipulate consumer reviews by: - Withholding high-quality information and disclosing low-quality information when consumers are naïve - Disclosing all information when consumers are sophisticated - Withholding low-quality information in advance when the market contains both naïve and sophisticated consumers

Appendix B: Variables Measured in the Pretest

Assume that you were considering purchasing the camera. Using the scales below, how would you describe the review above?

Extremity: (Lee et al., 2009)
- not at all positive / very positive
- not at all pleasant / very pleasant

(or)
- not at all negative / very negative
- not at all unpleasant / very unpleasant

Information quantity: (Gao et al., 2012)
- contains very little information / contains a great deal of information
- information contained in the review was not thorough at all / information contained in the review was very thorough

Concreteness: (Keller and Block, 1997)
- not at all concrete / very concrete
- not at all specific / very specific

Helpfulness: (Sen and Lerman, 2007)
- not at all helpful / very helpful
- not at all informative / very informative

Emotional intensity: (Jensen et al., 2013)
- contains little emotion / contains a great deal of emotion
- contains no feelings / contains a lot of feelings

Realism: (Mafael et al., 2016)
- not at all realistic / very realistic
- not at all real / very real

Reading difficulty: (Ermakova et al., 2014; Hall and Hanna, 2004)
- very hard to read / very easy to read
- very hard to understand / very easy to understand

Appendix C: Variables Measured in Studies 2 and 3

Purchase intention: (Dodds et al., 1991; Goldberg and Gorn, 1987) (Used in Studies 2 and 3)
Based on the information of Camera Model A/B, please answer the following questions.
- If you were thinking of buying a digital camera, how likely is it that you would buy Camera Model A/B?
- How likely is it that you would consider purchasing Camera Model A/B?
- How likely is it that Camera Model A/B would be a good choice for you?

Choice between two products: (Used in Studies 2 and 3)
Given a choice between the two cameras, which camera would you choose? (8-point scale)
- definitely choose Camera Model A / definitely choose Camera Model B

Average product rating:
Can you recall the average rating of Camera Model A/B based on hundreds of its prior customer reviews? (5-point scale)
- 3 stars / 3.5 stars / 4 stars / 4.5 stars / 5 stars (Used in Study 2)
- 2 stars / 2.5 stars / 3 stars / 3.5 stars / 4 stars (Used in Study 3)

Review valence: (MacKenzie and Lutz, 1989) (Used in Studies 2-3)
Using the scales below, overall, how would you describe the above 3 reviews as a whole?
- expresses very bad feelings about the camera / expresses very good feelings about the camera
- expresses very unfavorable feelings about the camera / expresses very favorable feelings about the camera
- expresses very unpleasant feelings about the camera / expresses very pleasant feelings about the camera

Concreteness: (Keller and Block, 1997) (Used in Study 3)
Using the scales below, overall, how would you describe the above review information as a whole?
- not at all concrete / very concrete
- not at all specific / very specific
- not at all vivid / very vivid
- not at all easy to imagine / very easy to imagine
- not at all easy to picture / very easy to picture
- not at all easy to relate to / very easy to relate to

REFERENCES

Ba, S., Jin, Y., Li, X., Lu, X. 2020. One size fits all? The differential impact of online reviews and coupons. *Production and Operations Management*. 29(10): 2403-2424.

Ceran, Y., Singh, H., Mookerjee, V. 2016. Knowing what your customer wants: Improving inventory allocation decisions in online movie rental systems. *Production and Operations Management*. 25(10): 1673-1688.

Chen, H., Zheng, Z., Ceran, Y. 2016. De-biasing the reporting bias in social media analytics. *Production and Operations Management*. 25(5): 849-865.

Dodds, W.B., Monroe, K.B., Grewal, D. 1991. Effects of price, brand, and store information on buyers' product evaluations. *Journal of Marketing Research*. 28(3): 307-319.

Ermakova, T., Baumann, A., Fabian, B., Krasnova, H. 2014. Privacy policies and users' trust: Does readability matter? *Proceedings of the 20th Americas Conference on Information Systems*. 1-12.

Gao, J., Zhang, C., Wang, K., Ba, S. 2012. Understanding online purchase decision making: The effects of unconscious thought, information quality, and information quantity. *Decision Support Systems*. 53(4): 772-781.

Goldberg, M.E., Gorn, G.J. 1987. Happy and sad TV programs: How they affect reactions to commercials. *Journal of Consumer Research*. 14(3): 387-403.

Gu, B., Ye, Q. 2014. First step in social media: Measuring the influence of online management responses on customer satisfaction. *Production and Operations Management*. 23(4): 570-582.

Guan, X., Wang, Y., Yi, Z., Chen, Y.J. 2020. Inducing consumer online reviews via disclosure. *Production and Operations Management*. 29(8): 1956-1971.

Hall, R.H., Hanna, P. 2004. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & Information Technology*. 23(3): 183-195.

Jensen, M.L., Averbeck, J.M., Zhang, Z., Wright, K.B. 2013. Credibility of anonymous online product reviews: A language expectancy perspective. *Journal of Management Information Systems*. 30(1): 293-324.

Keller, P.A., Block, L.G. 1997. Vividness effects: A resource-matching perspective. *Journal of Consumer Research*. 24(3): 295-304.

Lau, R.Y.K., Zhang, W., Xu, W. 2018. Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Production and Operations Management*. 27(10): 1775-1794.

Lee, M., Rodgers, S., Kim, M. 2009. Effects of valence and extremity of eWOM on attitude toward the brand and website. *Journal of Current Issues & Research in Advertising*. 31(2): 1-11.

MacKenzie, S.B., Lutz, R.J. 1989. An empirical examination of the structural antecedents of attitude toward the ad in an advertising pretesting context. *The Journal of Marketing*. 53(2): 48-65.

Mafael, A., Gottschalk, S.A., Kreis, H. 2016. Examining biased assimilation of brand-related online reviews. *Journal of Interactive Marketing*. 36: 91-106.

Sen, S., Lerman, D. 2007. Why are you telling me this? An examination into negative consumer reviews on the web. *Journal of Interactive Marketing*. 21(4): 76-94.

Sun, H., Xu, L. 2018. Online reviews and collaborative service provision: A signal-jamming model. *Production and Operations Management*. 27(11): 1960-1977.

Wang, Y., Goes, P., Wei, Z., Zeng, D. 2019. Production of online word-of-mouth: Peer effects and the moderation of user characteristics. *Production and Operations Management*. 28(7): 1621-1640.

Yan, L., Yan, X., Tan, Y., Sun, S.X. 2019. Shared minds: How patients use collaborative information sharing via social media platforms. *Production and Operations Management*. 28(1): 9-26.