THE UNIVERSITY OF CHICAGO

HOW PRODUCT REVIEWS IMPACT CONSUMERS' JUDGMENTS, EMOTIONS, AND PURCHASE BEHAVIORS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
DANIEL KATZ

CHICAGO, ILLINOIS
AUGUST 2024

Table of Contents

List of Tables iv

List of Figures v

Acknowledgments vi

Overview 1

Chapter 1 – Confirmation Bias in the Perceived Helpfulness of Product Reviews 3

- Abstract 4
- Theoretical Background 5
- Overview of Current Research 12
- Studies 1A & 1B: Mean Ratings and Judgments of Review Helpfulness 14
- Study 2: The Impact of Summary Information on Review Helpfulness 19
- Study 3: Belief Updating 25
- Study 4: Sentence-Level Helpfulness 30
- Study 5: Search for Reviews 34
- General Discussion 39

Chapter 2 – Beyond Unidimensional Sentiment Analysis: The Effect of Valence and Arousal on Evoked Emotion and Product Evaluations 46

- Abstract 47
- Theoretical Background 51
- Overview of Current Research 59
- Overview of Studies 63
- Study 1: The Importance of Evoked Emotion 65
- Study 2: Measuring Valence of Mixed Reviews 70
- Study 3: The Effect of Arousal on Purchase Behavior 75
- Study 4A: Discrete Emotional States and Purchase Behavior 82
- Study 4B: Discrete Emotional States and Purchase Behavior 87
- Study 5: The Impact of Basic Emotions on Valence, Arousal, and Product Evaluations 92
- General Discussion 98

Supplemental Material for Chapter 1 106

- Appendix 1.A: Supplemental Study 1 – Absolute Deviation and Willingness to Pay 106

Appendix 1.B: Supplemental Study 2 – Direction of Confirmation Bias 108

Appendix 1.C: Cue Weighting from Study 3 114

Appendix 1.D: Comparing the Book and Blender 115

Appendix 1.E: Additional Analyses 118

Appendix 1.F: Norming Study for Reviews 123

Supplemental Material for Chapter 2 127

Appendix 2.A: Supplemental Study 1 – Intensity, Evoked Emotion, and Purchase Behavior 127

Appendix 2.B: Full Regression Tables and Mediation Results 135

Appendix 2.C: Study Stimuli 143

References.. 152

List of Tables

Table 1.1: Summary of Studies and Hypotheses 14
Table 1.2: Study 1 Helpfulness Regression 18
Table 1.3: Study 2 Helpfulness Regression 23
Table 1.4: Study 3 Helpfulness Regression 28
Table 1.5: Review for Study 4 31
Table 1.6: Study 4 Mediation Results 33
Table 1.7: Study 5 Search Regression 36
Table 1.8: Study 5 Helpfulness Regression 37
Table 2.1: Dimensional Models of Emotion 52
Table 2.2: Categorical Models of Emotion 58
Table 2.3: Summary of Study Results 65
Table 2.4: Study 1 Regression Results 69
Table 2.5: Study 2 Sample Stimuli 72
Table 2.6: Study 3 Sample Stimuli 77
Table 2.7: Study 4B Emotional States 88
Table 2.8: Study 5 Sample Stimuli 94

List of Figures

Figure 1.1: Study 1A Helpfulness Ratings.......... 16
Figure 1.2: Study 1B Helpfulness Ratings.......... 17
Figure 1.3: Sample Distributions from Study 2 20
Figure 1.4: Study 2 Helpfulness Ratings.......... 22
Figure 1.5: Study 3 Helpfulness Ratings.......... 27
Figure 1.6: Study 3 Model Predictions for Attitude Change 29
Figure 1.7: Study 4 Helpfulness Ratings.......... 32
Figure 1.8: Study 4 Moderated Mediation Model 33
Figure 1.9: Study 5 Search Behavior and Helpfulness Ratings.......... 38
Figure 2.1: Emotional Wheel.......... 59
Figure 2.2: Theoretical Model for Evoked Emotion 61
Figure 2.3: Theoretical Model for Perceived Quality and WTP.......... 63
Figure 2.4: Study 2 Quality Judgments 73
Figure 2.5: Study 3 Results 81
Figure 2.6: Study 4A Results.......... 86
Figure 2.7: Study 4B Results.......... 91
Figure 2.8: Study 5 Results 97

Acknowledgments

I am grateful beyond words for the help and guidance from my co-chairs Dan Bartels and Abby Sussman. They have mentored me throughout my PhD and molded me into the researcher I am today. Their contributions to this dissertation and my development as a whole have been invaluable. I would also like to thank the other members of my committee, Oleg Urminsky and Reid Hastie, for their crucial suggestions regarding the direction of both chapters. Lastly, I received tremendous help from all the Chicago Booth CDR faculty and PhD students throughout my dissertation work.

I would also like to thank my family for their support. My mom and brothers were always there for me if I needed someone to proofread something quickly or test how long a survey would take. They kept me going throughout the program and I definitely could not have completed it without them.

Overview

This dissertation explores the effect product reviews have on consumer emotions, judgments, and decisions. Especially with the continued growth of e-commerce, product reviews play an important role in consumer decisions. More than 90% of consumers in the United States have used online reviews to help them make a purchase (Kaemingk, 2020). The majority of consumers trust these online reviews at least as much as they trust personal and expert recommendations (Galante, 2018; Statista, 2021). Given the extensive use of review information, it is critical for marketers and researchers to understand how these reviews impact consumers.

Chapter 1 examines how summary information about reviews for a product influences perceptions of individual reviews. Specifically, we study how manipulating the mean rating influences subsequent judgments of review helpfulness and search behavior. We find evidence of confirmation bias (Nickerson, 1998). Reviews that have ratings close to or at the mean (i.e., confirmed the mean) are rated as more helpful, lead to more extreme belief updating, and are more likely to be searched than reviews with ratings further from the mean. We also find process evidence that suggests the mean rating significantly influences how consumers weight the information in reviews, with greater weight being placed on information that confirms the mean rating. Lastly, we find participants are more likely to search for reviews near the mean when they could freely select which reviews to read. Taken together, these results suggest there is significant confirmation bias in consumers' judgments and behaviors when they are exposed to a product's mean rating.

Chapter 2 examines the role of emotion in product reviews and the effect it has on purchase behavior. Consumers consider the content or text of a review to be a highly influential feature of online reviews, above and beyond star ratings and total number of reviews (Podium, 2017). Additionally, sentiment analysis tools have surged in popularity, especially in marketing (e.g., https://www.revuze.it/). However, these tools often provide a simplistic view of the emotional content and how it may impact consumers. Thus, Chapter 2 studies how the emotional content of a review influences the emotions experienced by consumers as they read the review, as well as their eventual product evaluations. First, we find the emotion experienced by the consumer reading the review to be a stronger predictor of perceived product quality than the emotion expressed by the author of the review. Second, we demonstrate the need to measure positive and negative emotion on separate scales, as opposed to treating them as opposite ends of a single scale. When measured with a single scale, it is ambiguous whether the midpoint refers to a review that is fairly bland or one with a high degree of conflicting emotions. Lastly, we establish the need to consider arousal as an additional dimension when measuring emotions in reviews. Valence and arousal jointly impact product evaluations by influencing the amount of positive and negative emotion felt by the reader. These results highlight the advantage of going beyond a single, unidimensional scale when measuring emotion in product reviews.

Chapter 1 – Confirmation Bias in the Perceived Helpfulness of Product Reviews

Abstract

Online reviews can significantly influence consumer behavior. While there has been a large amount of work trying to estimate the impact of online reviews on sales (see Floyd et. al., 2014 for a meta-analysis), there is less work on what makes consumers perceive a review as helpful when making a purchase decision. We investigated the relationship between a review's perceived helpfulness and how much the review's rating deviates from the product's mean rating. We also studied the effect of a review's absolute deviation from the mean rating on how consumers acquire information and update their beliefs. Reviews that were close to the mean were rated as more helpful, led to more extreme belief updating, and were more likely to be searched, relative to reviews that deviated further from the mean. We found evidence these judgments and behaviors are due to confirmation bias.

In an increasingly digital economy, understanding the way the online marketplace impacts consumers is of great importance. Online reviews can strongly influence consumer behavior. More than 90% of consumers in the United States have used online reviews to help them make a purchase (Kaemingk, 2020). While there has been a large amount of work trying to estimate the impact of online reviews on sales (see Floyd et. al., 2014 for a meta-analysis), less work explores what makes reviews helpful for making a purchase decision. We investigated how a review's absolute deviation from the product's mean rating affects perceived helpfulness, as well as belief updating and consumer search.

Theoretical Background

Mean Ratings and Review Helpfulness

There is a sizeable literature on review helpfulness, but there is relatively little on the impact of summary information. Most previous work uses observational data, so it is difficult to make causal claims about the impact of summary information on helpfulness. For example, prior work has found that, for Amazon book reviews, there is a negative correlation between the number of helpful votes a review receives and its absolute deviation from the mean rating (Danescu-Niculescu-Mizil et. al., 2009; Bau & Chau, 2016). However, the correlation between these variables does not reveal whether the reviews' absolute deviations from the mean caused differences in the number of helpful votes. The correlational data is also mixed, as one dataset shows a positive correlation between helpfulness and absolute deviation from the mean for reviews of home goods products (Kupor & Tormala, 2018).

There is little experimental work that directly manipulates and measures the causal impact of the absolute deviation from the mean on helpfulness, and previous research yields mixed findings. One paper uses a natural experiment with the Apple App Store to demonstrate a causal, negative relationship between absolute deviation from the mean and review helpfulness (Yin, Mitra, & Zhang, 2016). The authors claim this is evidence of confirmation bias, but the observational data provide no process evidence to support the claim that confirmation bias is the mechanism driving their results. Relatively few lab studies probe the cognitive processes responsible for this relationship, and their results are mixed. In one study, Kupor and Tormala (2018) found reviews that deviate from the mean were more helpful because the mean rating can create a social default, leading reviews that deviate from the mean to be perceived as more thoughtful. On the other hand, there are lab experiments that find reviews are less helpful as they stray further from the mean because the reviewer is viewed as less credible (Qiu, Pang, & Lin, 2012). Our research sheds light on the underlying processes responsible for the effect of absolute deviation from the mean on perceived helpfulness.

The causal impact of the mean rating on review helpfulness is an open question, and the literature has not fully explored the cognitive processes that would lead absolute deviation from the mean to impact helpfulness. The current studies aim to inform these open questions, asking whether and why absolute deviation from the mean contributes to perceived helpfulness. In addition to testing this relationship between absolute deviation from the mean rating and helpfulness, we sought to uncover how this pattern arises and what its behavioral consequences are. To do so, we drew upon several streams of research involving heuristics and biases and how they relate consumer judgments and behaviors. We motivate several predictions for how the mean product rating might impact review helpfulness based on previous literature.

One well-studied behavioral tendency that could be relevant to the relationship between product ratings and review helpfulness is confirmation bias. Confirmation bias is an overarching term that refers to multiple psychological tendencies to search for, interpret, and recall information in a way that is consistent with one's own beliefs, expectations, or hypotheses (Klayman, 1995; Nickerson, 1998). With this broad definition, it's clear that there is more than one kind of confirmation bias. This relates to review helpfulness because these confirmation biases suggest people will perceive reviews to be more helpful when they are consistent with preexisting information, including knowledge of the product's average rating.

Of course, what causes a review to be helpful may depend on the goal of the consumer reading it, as confirmation bias can lead consumers to interpret information in a way that is congruent with their goals (Klayman, 1995). For example, a consumer may be looking at positive reviews to justify a purchase they want to make but is too expensive. Although goals are important for purchase decisions and confirmation bias, we abstract away from them in this paper due to the multitude of goals consumers could have. We provided participants with a generic goal of deciding which reviews would be most helpful if they were deciding whether or not to buy the specific product in the review, telling them they are already shopping for that type of product. In this case, confirmation bias would predict a negative relationship between helpfulness and absolute deviation from the mean.

H1a: Reviews closest to the mean will be judged as most helpful.

Despite this research on confirmation biases, the opposite prediction is plausible as well. One reason for this is reviews further from the mean could have a greater influence on one's expected level of quality. Some normative models of information search prescribe this behavior (e.g., Shannon, 1948), as the reviews far from the mean may contain more unique information that is not well-represented by the mean. Reviews close to the mean may often have information that is relatively more redundant when the mean is already known. If one adopts the mean rating as an initial expectation for product quality, this theory predicts reviews far from the mean will cause the greatest change in expected quality.

H1b: Reviews furthest from the mean will be judged as most helpful.

Another robust heuristic that could affect the relationship between a review's helpfulness and its absolute deviation from the mean is representativeness-based reasoning (Tversky & Kahneman, 1973). The most well-known case of this is one where participants read about a person named Linda who majored in philosophy and is concerned with discrimination and social justice. When asked, many participants judge that it is more likely that Linda is a feminist bank teller than a bank teller, which is impossible, as the latter category subsumes the former. Several investigations have followed up by exploring the pervasiveness of the representativeness heuristic, characterizing it in terms of formal models, and exploring when relying on this heuristic leads to better or worse outcomes (Tenenbaum & Griffiths, 2001; Bhatia, 2015). In the context of product reviews, consumers may find reviews helpful if they think the review represents a highly typical consumer experience. We consider the modal rating to be representative of a typical consumer, as it is the most common rating. If consumers engage in representative-based thinking, they might find representative reviews to be most helpful. This hypothesis differs from H1a because the mean rating of a product may not be the most representative rating. For example, a product containing only one- and five-star reviews could have a mean rating of three stars, even if no three-star reviews exist. We manipulated the distribution of review ratings to differentiate between these two hypotheses.

H1c: Reviews closest to the mode will be judged as most helpful.

As all these predictions are backed by theory, and prior work on this topic shows mixed and/or correlational evidence, we were agnostic as to which of these hypotheses would be true. In our studies, we find support for H1a and do not find support for H1b or H1c. Thus, we will focus the rest of this section on confirmation bias and the role it plays in consumers' use of product reviews.

Confirmation Biases and Product Reviews

Confirmation biases can play a significant role in consumers' use of product reviews. There are two broad categories of confirmation bias: backward-looking (e.g., reinterpretation, biased decision weights, etc.) and forward-looking (e.g., information search, biased attention, etc.). Backward-looking biases impact how information is encoded and interpreted (Klayman, 1995). People may interpret evidence in a way that favors initial beliefs, independent of the information they seek out (Nickerson, 1998). This often results in consumers being too conservative when updating their beliefs (Dave & Wolfe, 2003). If consumers simply take the mean rating as an informed prior and then engage in Bayesian updating while reading reviews, then the amount by which they update their beliefs should be sensitive to how strong of a signal the mean is. Thus, belief updating from a single review should become less extreme as the total number of reviews increases, because it would strengthen the informed prior. However, if consumers are using the mean as a belief they need to confirm, we would not necessarily expect to see much sensitivity to the number of reviews that generated the mean rating. While this is a null hypothesis, it would be surprising if the effect of the mean rating on consumers' perceptions of helpfulness does not vary with the total number of reviews because that directly changes how reliable of a signal the mean rating is. Additionally, prior research finds that purchase likelihood tends to increase with the total number of reviews, a phenomenon dubbed "popularity bias," suggesting it is something consumers are sensitive to (Powell et. al., 2017; Heck, Seiling, & Bröder, 2020).

H2: The total number of reviews will not moderate the effect a review's absolute deviation from the mean on its perceived helpfulness.

Another important way confirmation bias affects belief updating is by leading consumers to selectively encode confirming and disconfirming information. For example, consumers tend to distort information in ways that favor their preferred brands (Russo, Meloy, & Medvec, 1998). This stems partly from consumers' desires to achieve consistency between old and new information (Russo et. al., 2008). For example, a consumer who initially believes a product is high quality may give deference to positive signals and dismiss negative signals. For similar reasons, consumers may give greater weight to evidence that confirms a hypothesis or initial belief, particularly if the evidence is ambiguous (Klayman, 1995).

H3: Participants will update their beliefs about a product more after reading reviews close to the mean than after reading reviews far from the mean.

H4: In mixed reviews, the positive (negative) information will be relatively more helpful when the mean rating is high (low).

Forward-looking confirmation bias affects how people acquire information (Klayman, 1995). People tend to explore options that are expected to confirm some preconception of interest, otherwise known as a positive-test strategy. This relates to the current research because people often observe summary information, like the product's mean rating, before reading reviews. If people encode that information as a relevant property of the product, they may disproportionately seek out reviews that are consistent with the mean. Note, we make no claims about whether this is a mistake on the part of the consumer.

H5: Reviews close to the mean are more likely to be searched.

Overview of Current Research

In this work, we explored the relationship between a review's absolute deviation from the mean rating and its helpfulness. Five studies show that a review's helpfulness declines as the absolute difference between the review's rating and the mean product rating increases. Further, we attempted to understand why absolute deviation from the mean rating impacts review helpfulness and what the consequences are. All studies focused on the effect of absolute deviation due to findings in prior literature (e.g., Bau & Chau, 2016; Kupor & Tormala, 2018) and because the direction of the deviation did not qualitatively influence the results (see Appendix 1.E for analysis using signed deviation from the mean).

In Studies 1A and 1B, we examined the effect of absolute deviation from the mean on review helpfulness, providing an initial investigation of confirmation bias in judgments of review helpfulness. Study 2 manipulated several other types of summary information to study its impact on helpfulness. The mean rating was the only factor that significantly affected helpfulness. Studies 3 through 5 documented psychological and behavioral patterns consistent with confirmation bias. In Study 3, we measured how participants update their beliefs while reading a set of reviews. Study 4 examined how the mean rating influences what parts of a review are most helpful. Lastly, in Study 5, we observed how the mean rating affected which reviews participants chose to read. These five studies reveal a causal, negative impact of a review's absolute deviation from the mean on its perceived helpfulness. They also provide a clearer understanding of the processes underlying this relationship and explain some downstream consequences. Studies 1A, 1B, 2, 4, and 5 were preregistered. All preregistrations and full study materials are available on OSF[1] at: https://osf.io/kum3b/?view_only=1ce7516de70b4cd294bd4b83f2286e30

The studies in this paper make significant contributions to the existing literature. First, it provides causal evidence that a review's helpfulness depends on its absolute deviation from the mean rating. While there is prior research supporting that claim, much of it is correlational. Additionally, both the correlational and causal findings in prior research have been mixed. Our findings strengthen the claim that there is a negative relationship between a review's helpfulness and its absolute deviation from the mean. We also provide the first process level evidence that this relationship is driven, at least in part, by confirmation bias. Lastly, we show how this process affects downstream search behavior. Via these contributions, our work augments the field's understanding of how mean ratings affect review helpfulness and consumer behavior.

[1] This analyses in the paper may be slightly different from the main analyses we pre-registered, but they yield the same qualitative results. Any additional preregistered analyses are in Appendix 1.E.

Table 1.1: Summary of Studies and Hypotheses

Study	Main Hypotheses Tested	Results
1A and 1B	H1a: Reviews near mean are more helpful H1b: Reviews far from mean are more helpful	Supports H1a Does not support H1b
2	H1c: Reviews near mode are more helpful H2: Total number of reviews does not affect helpfulness	Does not support H1c Supports H2
3	H3: Reviews near mean cause more belief updating	Supports H3
4	H4: In mixed reviews, positive (negative) information is more helpful when the mean is high (low)	Supports H4
5	H5: Reviews near the mean are more likely to be searched	Supports H5

Studies 1A & 1B: Mean Ratings and Judgments of Review Helpfulness

Previous literature presents conflicting findings on whether reviews close to the mean are more or less helpful than reviews far from the mean (Danescu-Niculescu-Mizil et. al., 2009; Qiu, Pang, & Lin, 2012; Bau & Chau, 2016; Yin, Mitra, & Zhang, 2016; Kupor & Tormala, 2018). In Studies 1A and 1B, we investigated how a review's absolute deviation from the mean rating influenced judgments of how helpful it is. We manipulated absolute deviation from the mean in two different ways to help ensure the results (and results in prior research) are not an artifact of the environment.

Experimental Design and Procedure

Study 1A used a 2(review rating: 2 or 4) x 2(mean rating: 3 or Equal to Review Rating) between-subjects design. Participants saw only one review and answered the following question: “How helpful would this review be when deciding whether to buy this book?” (1 = Not helpful at all, 7 = Very helpful). Our intention was to examine whether reviews are more helpful when they are equal to the mean rating. The text of both the two- and four-star reviews were held constant across mean rating conditions, so any difference in helpfulness can be attributed to the manipulation of the mean rating. Note, we used the three-star review as a comparison so both the two- and four-star reviews were tested against the same review, and because the mean cannot reasonably be one or five. We separated the two- and four-star means when comparing to the three-star mean in case there is a difference when comparing the three-star mean to positive versus negative reviews.

Study 1B used a 2(mean rating: 2 or 4) x 5(star rating of review: 1, 2, 3, 4, or 5) between-subjects design. Participants saw only one review and answered the same helpfulness question as in Study 1A. We used reviews adapted from Amazon book reviews. The reviews we used were normed in a prior study to be reviews that participants saw as highly typical of one particular star rating and atypical of other star ratings[2]. The advantage of this design is we do not expect participants to perceive any mismatch between the text of a review and its star rating.

Participants

In Study 1A, 600 participants completed the survey on Prolific (M_{age} = 36, 48% female). Six participants were excluded due to a memory check failure, leaving 594 valid completions.

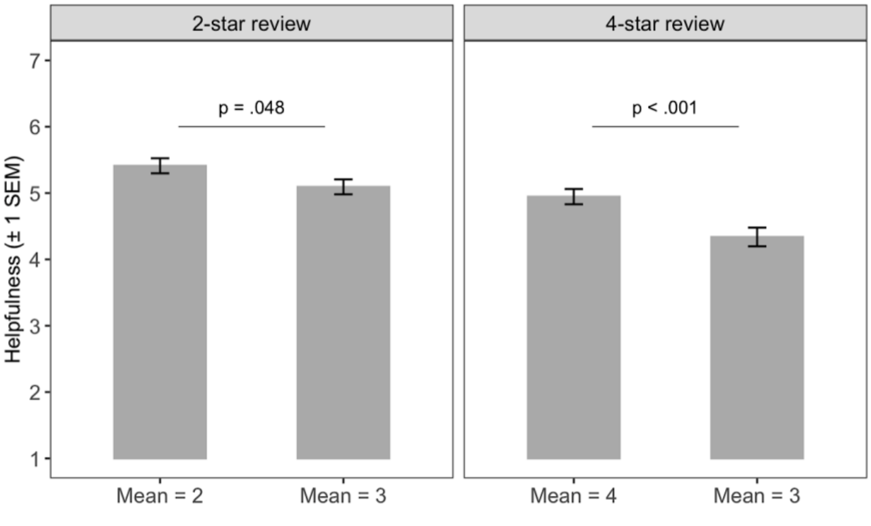
[2] The reviews from this norming study were also used for the remaining studies. See Appendix 1.F for details.

In Study 1B, 296 participants completed the survey on Prolific (M_{age} = 34, 46% female). Two participants were excluded due to a memory check failure, leaving 294 valid completions.

Results

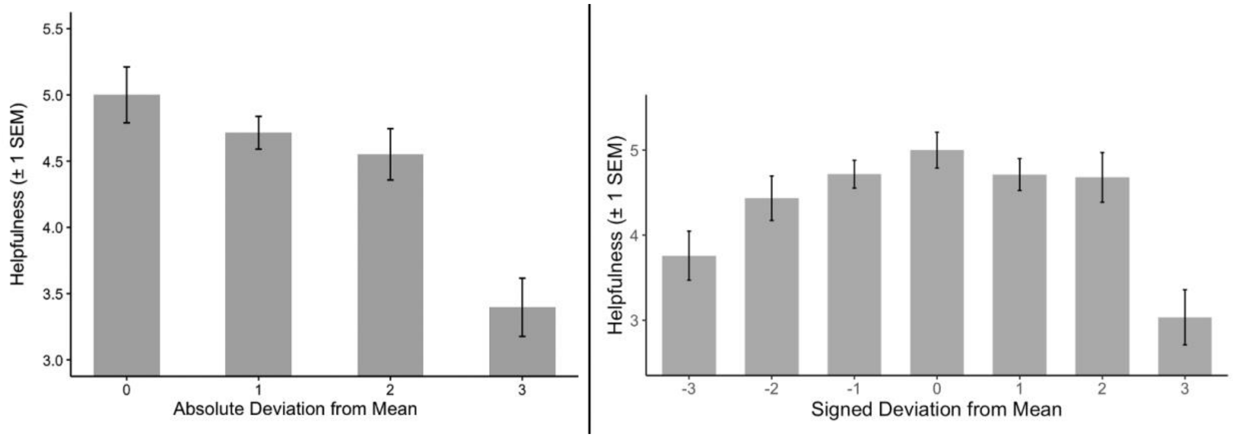
In Study 1A, we observed that reviews were seen as being more helpful when the review rating equaled the mean rating. For two-star reviews, helpfulness was higher when the mean was two than when it was three ($M_{Mean=2}$ = 5.41, $M_{Mean=3}$ = 5.09, $SD_{Mean=2}$ = 1.39, $SD_{Mean=3}$ = 1.38, t(299) = 1.99, p = .048, Cohen's d = 0.23). Similarly, for four-star reviews, helpfulness was higher when the mean was four than when it was three ($M_{Mean=4}$ = 4.95, $M_{Mean=3}$ = 4.34, $SD_{Mean=4}$ = 1.40, $SD_{Mean=3}$ = 1.70, t(279) = 3.34, p < .001, Cohen's d = 0.39).

Figure 1.1: Study 1A Helpfulness Ratings



In Study 1B, we observed that increasing the absolute deviation between a review's rating and the mean rating caused perceived helpfulness to decrease (Pearson's r = -0.32, Spearman's ρ = -0.31, p's < .001). In other words, participants rated reviews close to the mean as more helpful than reviews far from the mean (see Figure 1.2). Figure 1.2 also shows the relationship between helpfulness and signed deviation from the mean (i.e., not the absolute value). Across studies, the pattern is relatively symmetric regardless of whether the absolute deviation is above or below the mean. The important point is that, regardless of direction, helpfulness decreases as reviews stray further from the mean.

Figure 1.2: Study 1B Helpfulness Ratings



To test this relationship, we regressed review helpfulness on absolute deviation from the mean, the star rating of the review (centered at 3; i.e., it ranges from -2 to 2), and their interaction. We include star rating in the model because previous research on review helpfulness has found differences in helpfulness for positive versus negative reviews (e.g., Sen & Lerman, 2007). We standardized all variables, so the coefficients represent a measure of effect size on the same scale as a correlation coefficient (Peterson and Brown, 2005). We found a significant, negative coefficient for absolute deviation from the mean (see Table 1.2). There was no main effect or interaction with the star rating of the review participants saw.

Table 1.2: Study 1 Helpfulness Regression

Predictors	Estimate	t	p
Intercept	0.00	0.00	.998
Absolute Deviation from Mean	**-0.32**	**-5.79**	**< .001**
Star Rating	0.04	0.67	.504
Absolute Deviation from Mean*Star Rating	0.07	1.36	.176
N_{id}	294		
R^2	0.11		
R^2 adjusted	0.10		

NOTE – All variables standardized

Discussion

Studies 1A and 1B supports H1a, as reviews near the mean were judged as more helpful than reviews far from the mean. In Study 1A, the exact same review was seen as more helpful when its rating was consistent with the mean. This is consistent with confirmation bias. Study 1B shows this pattern exists across the full range of star ratings. The results do not support H1b, which predicts the opposite pattern. In these studies, we focused on manipulating the mean rating, holding other factors constant. In Study 2, we vary other aspects of summary information that could influence review helpfulness.

Study 2: The Impact of Summary Information on Review Helpfulness

The mean rating of a product is often just one of several pieces of summary information consumers have when shopping online. Many websites also display the distribution of star ratings a product has received as well as the total number of ratings. While we are motivated to explore and explain the relationship between the mean rating and review helpfulness, we believed these additional factors could also lead the summary information to influence perceived helpfulness.

Experimental Design and Procedure

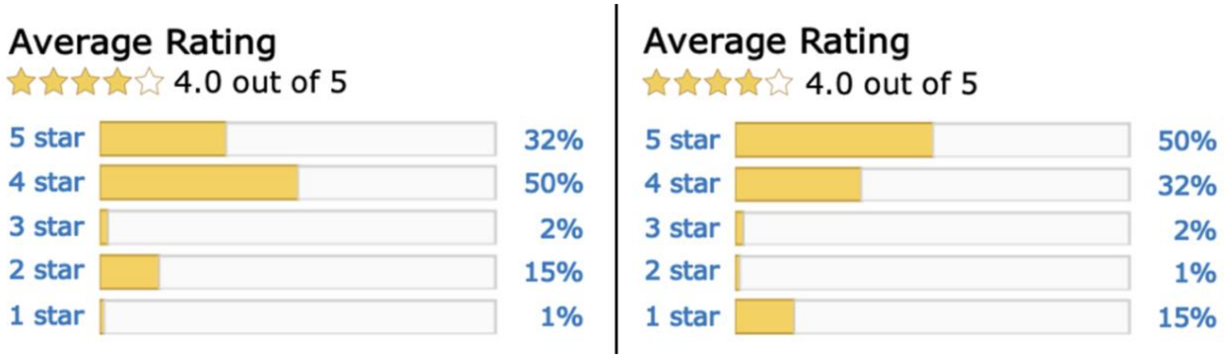
Study 2 used a 3(mean rating: 2, 3, or 4) x 2(distribution: mean = mode, mean ≠ mode) x 2(product: blender, book) x 2(total number of reviews: low = 84, high = 984)[3] x 5(star rating of review: 1, 2, 3, 4, 5) between-subjects design. Participants saw one review and rated its helpfulness on a 7-point scale. This design allowed us to test the effect of summary information on review helpfulness more broadly.

The distribution of reviews carries a lot of information about how consumers feel about a product. We chose to manipulate the mode as a potential alternative to the hypothesis involving the mean. If consumers reading the reviews are trying to gauge the experience they are most likely to have if they purchase the product, the mode seems like a sensible reference point they could adopt. The distributions we used were constructed specifically to hold constant, within each mean rating, the percentage of reviews that were four or five stars and

[3] These numbers for total reviews were chosen from prior norming questions about what participants believed was a low and high number of reviews for books/blenders. Responses were extremely similar for both products.

the percentage of reviews that were one or two stars. The reason for this was to control for people’s tendency to group four- and five-star reviews together as “good” and group one- and two-star reviews together as “bad” (known as “binary bias”; Fisher, Newman, & Dhar, 2018). Figure 1.3 shows this manipulation for an average rating of four stars.

Figure 1.3: Sample Distributions from Study 2



NOTE- left: mean = mode, right: mean ≠ mode

We used two different products in this study, a blender and a book. We used these products for two reasons. First, this could potentially reconcile conflicting findings in the literature on review helpfulness. Prior work has found a negative relationship between absolute deviation from the mean and number of helpful votes in Amazon reviews, but it used a narrow set of product categories, namely books, music, and games (Bao & Chau, 2016). Other work has found the opposite relationship using a dataset of home goods and accessories (Kupor & Tormala, 2018). Thus, we chose one product from each of these categories to see if product differences could help explain conflicting findings in past literature on the relationship between absolute deviation and helpfulness.

Second, these products differ on several key dimensions that prior literature suggests are important for review helpfulness, including how subjective the quality of the product is (i.e., the variance in quality judgments), whether it is more hedonic or pragmatic (i.e., “utilitarian”), and whether the product is material or experiential. Some research has found consumers rely on reviews more for material products than for experiential products (Dai, Chan, & Mogilner, 2020). Additionally, there has been research showing that negative reviews are more helpful than positive reviews (a pattern referred to as “negativity bias”) for hedonic goods, but not for pragmatic goods (Sen & Lerman, 2007).

We also manipulated whether the total number of reviews was small or large. Past research has found one’s product preferences are significantly influenced by the number of total reviews a product has, such that consumers are typically more likely to choose an option that has the higher number of total reviews (Powell et. al., 2017; Heck, Seiling, & Bröder, 2020). Additionally, the mean rating should create a stronger prior belief about product quality when the total number of reviews is higher. When the mean is a stronger signal, each individual review should be less helpful when arriving at a decision of whether to by the product. However, this effect need not exist if consumers are engaging in confirmation bias.

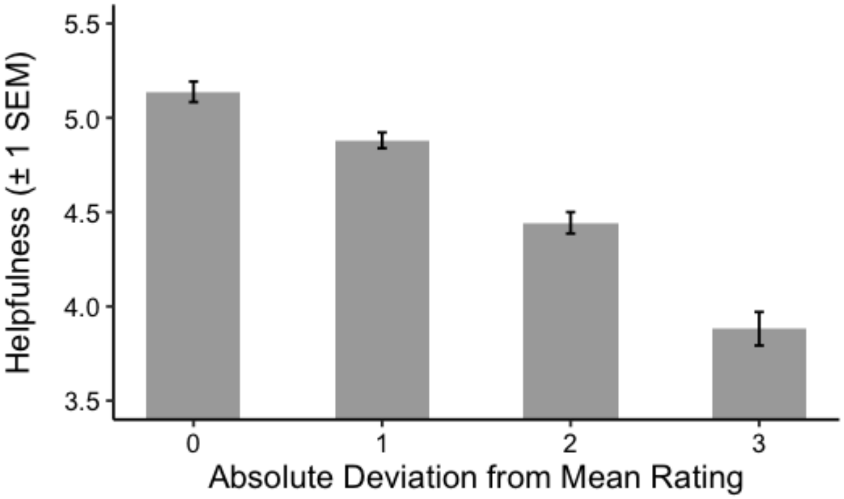
Participants

Three thousand, six hundred and eleven participants completed the survey on Prolific (M_{age} = 35, 52% female). Eight participants were excluded due to a memory check failure, leaving 3,603 valid completions.

Results

As in Studies 1A and 1B, reviews near the mean were rated as more helpful than reviews far from the mean. We observed a significant negative relationship between absolute deviation from the mean and rated helpfulness (Pearson's r = -0.23, Spearman's ρ = -0.21, p's < .001; see Figure 1.4).

Figure 1.4: Study 2 Helpfulness Ratings



We regressed review helpfulness on absolute deviation from the mean and the other experimental factors (the terms in the regression are in Table 1.3). We standardized all continuous variables and centered star rating at 3. We deviation coded all categorical factors so the coefficient for absolute deviation from the mean is an average across conditions[4]. We observed a significant negative relationship between a review's absolute deviation from the mean and its helpfulness rating.

[4] For example, Book takes a value of 0.5 and Blender takes a value of -0.5, as opposed to 1 and 0.

Table 1.3: Study 2 Helpfulness Regression

Predictors	Estimate	t	p
Intercept	-0.00	-0.02	.985
Absolute Deviation from Mean	**-0.23**	**-14.80**	**< .001**
Star Rating	-0.12	-7.04	< .001
Distribution [Mean ≠ Mode]	0.01	0.48	.633
Product [Book]	-0.37	-11.96	< .001
Total Reviews [High]	-0.01	-0.45	.654
Absolute Deviation from Mean*Star Rating	-0.04	-2.69	.007
Absolute Deviation from Mean*Distribution [Mean ≠ Mode]	0.04	1.27	.205
Absolute Deviation from Mean*Product [Book]	-0.01	-0.43	.668
Absolute Deviation from Mean*Total Reviews [High]	0.01	0.18	.854
Star Rating*Book	0.36	11.81	< .001
Observations	3603		
R^2	0.14		
R^2 adjusted	0.14		

NOTE – Continuous variables standardized; categorical variables deviation coded

Product type had a significant effect on helpfulness, but the effect of absolute deviation from the mean did not differ across products. We observe a significant interaction between product type and the star rating of the review, as expected. For the blender, we observed that negative reviews were rated as more helpful than positive reviews (i.e., “negativity bias”), as the coefficient on “Star Rating” is negative and the blender is the reference product. Conversely, participants who saw the book found positive reviews were more helpful (i.e., “positivity bias”), as the sum of the coefficients for “Star Rating” and “Star Rating*Book” is positive (-0.12 + 0.36 = 0.24). This pattern is consistent with previous research (Sen & Lerman, 2007). Neither the distribution of reviews nor the total number of reviews significantly influenced helpfulness judgments. Furthermore, the effect of absolute deviation from the mean on helpfulness did not vary across either of these factors. We did observe an interaction between a review's absolute deviation from the mean and its star rating, such that the effect is larger for reviews with lower ratings. However, that is inconsistent across studies and the effect size is dwarfed by the main effect of absolute deviation from the mean, which still has a negative relationship with helpfulness for all possible star ratings.

Discussion

Like Studies 1A and 1B, results from Study 2 supports H1a over H1b. The negative relationship between helpfulness and absolute deviation from the mean is most consistent with confirmation bias. Study 2 also tested H1c by manipulating the mode of the distribution, independently from the mean. Participants were not sensitive to differences in the distribution, suggesting the most helpful reviews were not the ones that were most representative of a typical consumer. However, we do continue to vary the distribution in this way in future studies for the purpose of stimulus sampling.

Supporting H2, participants' helpfulness judgments were insensitive to the total number of reviews, a pattern that is consistent with confirmation bias and inconsistent with Bayesian reasoning. This pattern suggests that participants may be trying to confirm the mean rather than simply using it as an informed prior. Future studies will reveal other patterns of judgments and behaviors that provide evidence of confirmation bias. Note, the "low" and "high" number of reviews we used came from a norming study where we asked participants what they think would be a small or large total. Thus, we do not believe the insensitivity is an artifact of the specific totals we used in our stimuli.

Regarding product type, our results are consistent with prior literature on negativity bias in reviews. We wanted to test whether the negative relationship between helpfulness and absolute deviation held in the presence of negativity bias and the opposite, positivity bias, which it did. Future studies will further investigate this difference between the two products.

Studies 1A, 1B, and 2 document a negative relationship between the helpfulness of a review and its absolute deviation from the mean rating. A similar correlation has been observed in several other papers (Danescu-Niculescu-Mizil et. al., 2009; Bau & Chau, 2016). Some prior work postulates this is a form of confirmation bias. However, testing that claim requires data that explores the underlying psychological processes, which has not been done. Studies 3 – 5 directly test this confirmation bias hypothesis by presenting evidence of the cognitive processes involved.

Study 3: Belief Updating

As consumers read through reviews, they update their beliefs about the product. It is plausible that the reviews consumers perceive as most helpful will also cause greater changes in beliefs about the product, while unhelpful reviews may have less of an impact. Study 3 investigates this possibility by measuring the degree to which participants update their beliefs about a product after reading each review. In the presence of confirmation bias, one would update their beliefs significantly after acquiring information that confirms a hypothesis, while there would be relatively less belief change after acquiring information that contradicts a hypothesis (Dave & Wolfe, 2003).

Experimental Design and Procedure

This study used a 3(mean rating: 2, 3, 4; between) x 2(distribution: mean = mode, mean ≠ mode; between) x 2(product: book, blender; between) x 5(star rating of reviews: 1, 2, 3, 4, 5; within) mixed design. Participants saw one review from each star rating in a random order. For each review, after providing their helpfulness judgments, participants answered the following question: "After reading the review, do you feel more positive or negative about the product?" (-3 = much more negative, 0 = no change, 3 = much more positive). This allowed us to measure the how much participants' attitudes toward the product changed after reading each review.[5]

Participants

Six hundred and thirty-two participants completed the survey on Prolific (M_{age} = 32, 51% female). Two participants were excluded due to a memory check failure, leaving 630 valid completions.

[5] We also asked how important the text of the review (relative to its star rating) was to their judgment of helpfulness. See Appendix 1.D for further details.

Results

Study 3's findings are consistent with earlier studies. Helpfulness was negatively related to the absolute deviation from the mean (Pearson's r = -0.17, Spearman's ρ = -0.18, p's < .001, see Figure 1.5). We ran the same regression from earlier studies, clustering standard errors by participant, and found a significant, negative effect of absolute deviation on review helpfulness (see Table 1.4).

Figure 1.5: Study 3 Helpfulness Ratings

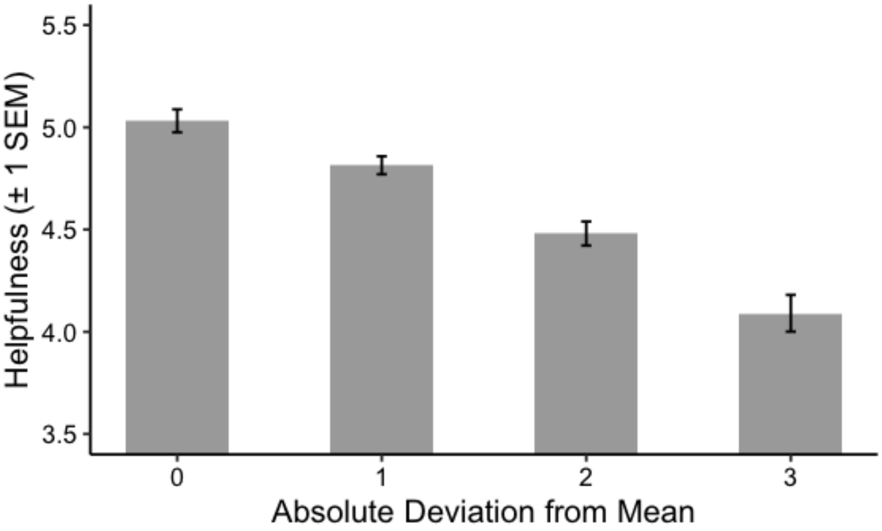


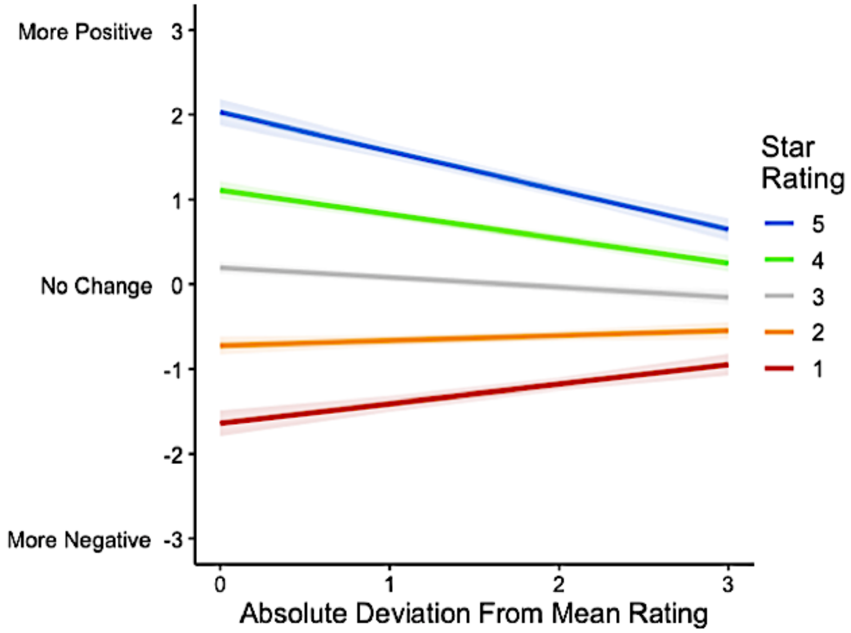
Table 1.4: Study 3 Helpfulness Regression

Predictors	Estimate	t	p
Intercept	0.00	0.00	.999
Absolute Deviation from Mean	**-0.18**	**-11.63**	**< .001**
Star Rating	-0.11	-5.50	< .001
Distribution [Mean ≠ Mode]	0.00	0.03	.977
Product [Book]	-0.37	-7.63	< .001
Absolute Deviation from Mean*Star Rating	-0.03	-1.45	.148
Absolute Deviation from Mean*Distribution [Mean ≠ Mode]	-0.01	-0.34	.731
Absolute Deviation from Mean*Product [Book]	-0.05	-1.70	.088
Star Rating*Product [Book]	0.51	15.60	< .001
Observations	3150		
R^2	0.15		
R^2 adjusted	0.15		

NOTE – Continuous variables standardized; categorical variables deviation coded

We ran the same regression with the belief updating measure as the dependent variable and found a significant interaction between absolute deviation from the mean and star rating of the review (Standardized $\beta_{\text{interaction}}$ = -0.14, t = -9.31, p < .001). This revealed that participants updated their beliefs about the product to a greater (lesser) degree after reading reviews close to (far from) the mean. This interaction can be more easily interpreted by examining the simple slopes. Figure 1.6 shows the effect of absolute deviation on attitude change at each star rating. The slopes for positive reviews are negative, indicating these reviews had more of a positive effect on product attitudes when they were close to the mean. Conversely, the slopes for negative reviews are positive, indicating these reviews had more of a negative effect on product attitudes when they were close to the mean (p < .001 for all simple slopes).

Figure 1.6: Study 3 Model Predictions for Attitude Change



Discussion

The results of Study 3 replicated the findings of prior studies and gave further insight into the processes underlying participants' judgements of review helpfulness. Supporting H3, participants updated their attitude toward the product to a greater degree for reviews close to the mean. This pattern suggests a form of backward-looking confirmation bias — participants adjusted their beliefs more when evidence confirmed the mean than when it did not. Note that we cannot determine, based on these data, if participants updated their beliefs more for confirmatory reviews because they were judged as more helpful or whether participants judged confirmatory reviews as more helpful because they updated their beliefs more. Further investigation is needed to identify the direction of causality between those measures. Regardless, Study 3 provides additional, direct evidence of confirmation bias.

Study 4: Sentence-Level Helpfulness

Because of confirmation bias, we expect consumers to fixate on positive information when the mean rating is high but fixate on negative information when the mean is low. This is a classic form of confirmation bias known as a positive-test strategy (Klayman, 1995). In one study, Shafir (1993) asked half the participants which of two parents they would award custody of a child to, the other half chose which parent they would deny custody of a child. In the former condition, participants looked for positive information that would support awarding custody. In the latter, participants looked for negative information that would support denying custody. We performed a similar study here to test for patterns of confirmation bias in review helpfulness judgments.

Experimental Design and Procedure

Study 4 used a 2(mean rating: 2 or 4) x 2(review rating: 2 or 4) between-subjects design. All participants saw a review with two positive and two negative sentences, in a random order. Additionally, the four- [two-] star reviews had a sentence stating the reviewer liked [disliked] the product (see Table 1.5). Participants read the review, then ranked how helpful each sentence was, and then rated the helpfulness of the review as a whole (just as in prior studies).

Table 1.5: Review for Study 4

Sentence 1:

I (dis)liked this book and would (not) recommend it.

Sentences 2-5 (randomized order):

The characters were developed well and I became invested in their story. I liked the ending, it tied everything together nicely. The plot was fairly predictable and there wasn't a lot of suspense. The writing wasn't always clear and I had to re-read some parts.

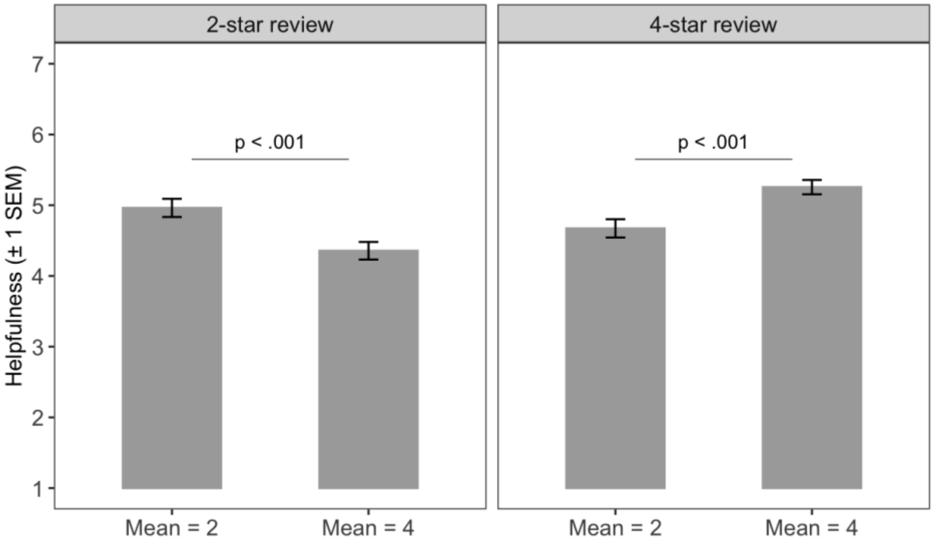
Participants

Six hundred and fifty-one participants completed the survey on Prolific (M_{age} = 40, 49% female). Six participants were excluded due to a memory check failure, leaving 645 valid completions.

Results

As in Study 1A, reviews were more helpful when they were consistent with the mean. When the mean was two, the two-star review was judged as being more helpful than the four-star review ($M_{Mean=2}$ = 4.96, $M_{Mean=4}$ = 4.67, $SD_{Mean=2}$ = 1.62, $SD_{Mean=4}$ = 1.64). The opposite was true when the mean was four ($M_{Mean=2}$ = 4.36, $M_{Mean=4}$ = 5.26, $SD_{Mean=2}$ = 1.58, $SD_{Mean=4}$ = 1.30; see Figure 1.7). When we regressed review helpfulness on mean rating, review rating, and their interaction, the interaction term was significant (Standardized β = 0.76, t(641) = 4.90, p < .001; see Appendix 1.E for full regression table).

Figure 1.7: Study 4 Helpfulness Ratings



Next, we examined how the mean rating affected which parts of the review participants found most helpful. We created a measure called "relative positivity," which is the mean rank of the negative sentences minus the mean rank of the positive sentences (note: a rank of one means it was the most helpful sentence). Only sentences 2-5 are included in this measure, as they are held constant across conditions (sentence 1 was also consistently rated as the least helpful). A t-test revealed that relative positivity was greater when the mean was four than when it was two ($M_{Mean=4}$ = 0.02, $M_{Mean=2}$ = -0.58, $SD_{Mean=4}$ = 1.71, $SD_{Mean=2}$ = 1.73, t(643) = 4.43; p < .001, Cohen's d = 0.35). In other words, the positive (negative) sentences in the review were seen as relatively more helpful when the mean was four (two). There was no interaction with review rating.

We then ran a moderated mediation analysis with mean rating as the independent variable, overall review helpfulness as the dependent variable, relative positivity as the mediator, and review rating as a moderator on the b- and c-paths (see Figure 1.8).[6] As expected, the indirect path for the two-star review was negative while the indirect for the four-star review was positive (though not significant). The index of moderated mediation (Hayes, 2015) was significant (Bootstrapped 95% CI = [0.04, 0.24]). See Appendix 1.E for all coefficients from regressions in the mediation model.

Figure 1.8: Study 4 Moderated Mediation Model

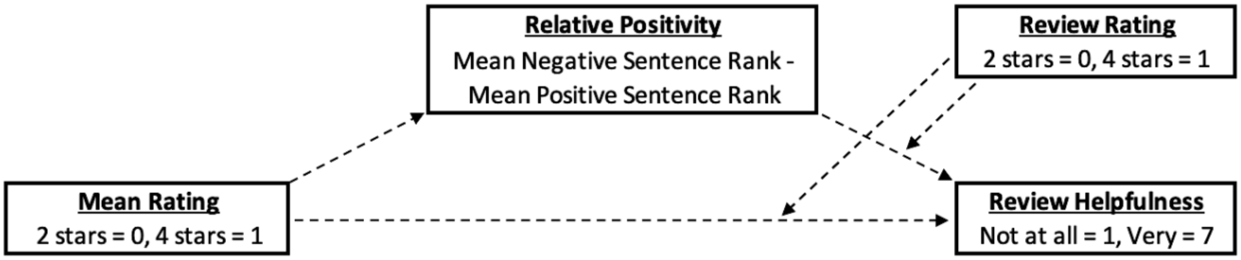


Table 1.6: Study 4 Mediation Results

	Estimate	95% Bootstrapped CI
Indirect Effect (2-star reviews)	-0.11	[-0.19, -0.04]
Indirect Effect (4-star reviews)	0.02	[-0.03, 0.08]
Index of Moderated Mediation	0.13	[0.04, 0.24]

Discussion

Study 4 finds additional evidence of confirmation bias in review helpfulness judgments. In addition to replicating our main result for the reviews' overall levels of helpfulness, we see

[6] We did not expect moderation on the a-path, and indeed there was no interaction between mean rating and review rating in a regression predicting relative positivity. Regardless of review rating, relative positivity was greater when the mean was four than when it was two. Appendix 1.E contains the output from a mediation model with moderation on all paths, which does not qualitatively impact the results.

significant differences in which aspects of the review are most helpful. Participants who saw a mean of four gave relatively more weight to the positive sentences, while participants who saw a mean of two gave relatively more weight to the negative sentences. This conceptually replicates a classic experiment in the confirmation bias literature within the context of consumer reviews to support the notion of confirmation bias as a mediating process for our prior findings. In Study 5, we sought to build on this by replicating another finding from the confirmation bias literature in our context, as well as demonstrate a behavioral consequence of that bias.

Study 5: Search for Reviews

Study 5 further examines the effect of the positive-test strategy heuristic (Klayman, 1995) on consumer behavior by studying search patterns. When searching for reviews to read, consumers are often first exposed to the mean product rating. In the presence of confirmation bias, we would expect participants to search more for information consistent with that prior piece of information. This suggests that a crucial behavioral consequence of this bias might be the way that consumers search for reviews.

Experimental Design and Procedure

This study used a 3(mean rating: 2, 3, or 4; between) x 2(distribution: mean = mode, mean ≠ mode; between) x 5(star rating of reviews that could be searched: 1, 2, 3, 4, 5; within) x 4(product: book, painting, blender, trash can; between) mixed design. The first (last) two products are hedonic (pragmatic), experiential (material), complex (simple), and have high (low) variance in quality judgments. These are dimensions along which the book and blender significantly differ and there is past literature suggesting these dimensions could affect the types of reviews participants find helpful (see Appendix 1.D for product ratings on these dimensions).

Unlike the previous studies, participants in Study 5 chose which reviews to read (as opposed to being randomly assigned or forced to respond to all five). Participants were required to search at least one review, after which they could terminate search at any time. The maximum number of reviews they could search was five (one from each star rating). Afterward, participants saw and gave helpfulness ratings for all five reviews, just as in prior studies.

Participants

Six hundred and four participants completed the survey on Prolific (M_{age} = 33, 52% female). Ten participants were excluded due to a memory check failure, leaving 600 valid completions.

Results

The mean rating participants saw significantly influenced their search behavior. For each participant, we calculated the average rating of all the reviews they searched (which could range from one to five). This measure reveals where participants were searching. Figure 1.9 shows a positive relationship between the mean rating of the reviews a participant chose to search and the product's mean rating (Pearson's r = 0.44, Spearman's ρ = 0.39, p's < .001). To formally test this relationship between search and the mean rating, we regressed the mean of reviews participants searched on the mean product rating and our other experimental factors (see Table 1.7). The results confirmed that the mean product rating had a significant, positive effect on the mean rating of the reviews a participant searched. In other words, participants were more likely to search for reviews close to the mean than for reviews far from the mean.

Table 1.7: Study 5 Search Regression

DV = Mean of Searched Reviews			
Predictors	Estimate	t	p
(Intercept)	0.00	0.00	.997
average rating	0.39	10.43	< .001
Mean ≠ Mode	0.13	1.71	.087
Hedonic Product	0.25	3.32	.001
Average Rating × Mean ≠ Mode	0.04	0.58	.564
Average Rating × Hedonic Product	0.06	0.79	.428
Observations	600		
R^2	.172		
R^2 adjusted	.165		

NOTE- continuous variables standardized

Additionally, as in previous studies, we found a negative relationship between a review's helpfulness and its absolute deviation from the mean (Pearson's r = -0.21, Spearman's ρ = -0.23, p's < .001; see Table 1.8 for full regression).

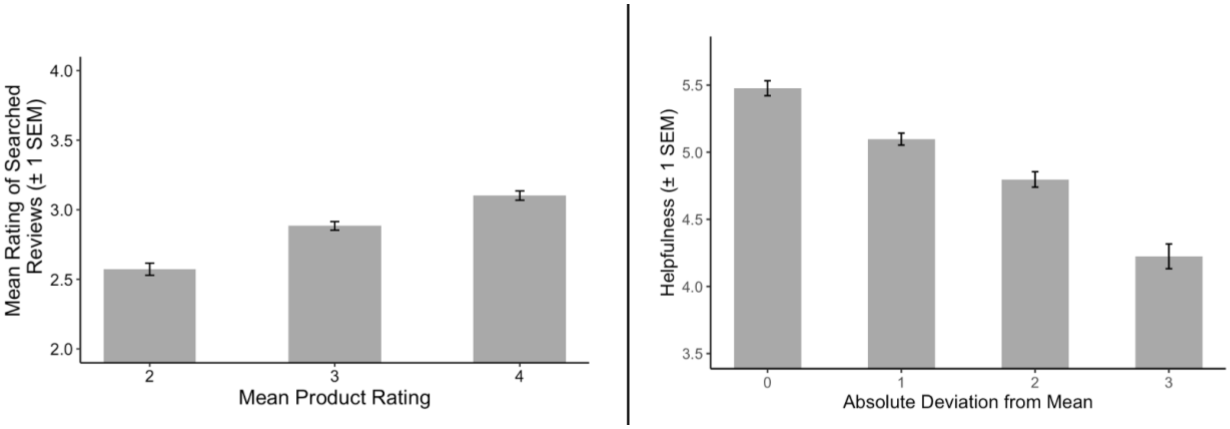
Table 1.8: Study 5 Helpfulness Regression

Helpfulness			
Predictors	Estimate	t	p
Intercept	0.00	0.01	.995
Absolute Deviation	-0.23	-12.55	< .001
Star Rating	-0.20	-8.91	< .001
Mean ≠ Mode	0.02	0.38	.701
Hedonic Product	-0.22	-4.89	< .001
Absolute Deviation*Star Rating	-0.03	-1.69	.092
Absolute Deviation*Mean ≠ Mode	-0.12	-3.35	.001
Absolute Deviation*Hedonic Product	-0.10	-2.59	.010
Star Rating*Hedonic Product	0.34	8.77	< .001
Observations	3000		
R^2	.146		
R^2 adjusted	.144		

NOTE- continuous variables standardized

Looking at the effect of product type, the results are similar to prior studies. For pragmatic products, like the blender, negative reviews were more helpful (i.e., the coefficient on “Star Rating” is negative). For hedonic products, like the book, positive reviews were more helpful (the sum of the “Star Rating” and “Star Rating*Hedonic Product” coefficients was positive; -0.20 + 0.34 = 0.14).

Figure 1.9: Study 5 Search Behavior and Helpfulness Ratings



NOTE – Search Behavior (left); Helpfulness Ratings (right)

Discussion

Results from this study replicate the pattern that review helpfulness is negatively related to absolute deviation from the mean. Additionally, we find this has significant consequences for the reviews participants chose to search. Participants tended to search for reviews that were close to the mean rating, exhibiting a positive-test strategy (Klayman, 1995). This is one way in which the effect of absolute deviation from the mean on review helpfulness can have a significant impact on consumer behavior. In this study, for the helpfulness regression only, there was an interaction between absolute deviation from the mean and the distribution (i.e., whether the mean = mode). We cannot explain why it happens to be significant here, but the clear consensus from the studies collectively is that the distribution did not affect perceived helpfulness (we only continued to vary it for the purpose of stimulus sampling).

We also revealed a predictable difference in whether positive or negative reviews will be more helpful based on the type of product. For pragmatic products, where perceptions of quality are likely to be more homogenous, negative reviews were more helpful. This is consistent with the negativity bias often found in the literature. However, for hedonic products, where perceptions of quality are more varied, positive reviews were more helpful. Product type also impacted search, as participants searched for more positive (negative) reviews for the hedonic (pragmatic) products. This interaction advances the field's understanding of the effect of valence on review helpfulness and search, particularly the moderating role of product type.

General Discussion

Five studies found a negative relationship between the perceived helpfulness of a review and its absolute deviation from the mean product rating. This pattern is consistent with research on confirmation bias. Additionally, Studies 3 – 5 provide more direct evidence of confirmation bias by investigating some of the cognitive processes responsible for this pattern. Study 3 shows participants updated their attitudes toward the product to a greater degree after reading reviews that were close to the mean. Study 4 shows participants gave more weight to aspects of a review that are consistent with the mean. Lastly, Study 5 presents a behavioral consequence of this bias by showing that participants chose to search for reviews consistent with the mean more than for reviews inconsistent with the mean. Together, these findings contribute to the literatures on consumer reviews, consumer search, and confirmation biases.

These studies show evidence of both backward- (Study 3) and forward-looking (Studies 4 and 5) confirmation bias. We ran an additional study where we manipulated the order in which participants saw the review and the summary information. We found a negative relationship between helpfulness and absolute deviation from the mean, even when the mean was presented after reading the review (but before the judgment). This suggests participants were using the mean when making retroactive judgments of review helpfulness, as the mean was unknown when reading the review. This provides further evidence that both classes of confirmation bias are present in this context (see Appendix 1.B for full study details).

Several factors we manipulated did not have an effect on review helpfulness. A review's absolute deviation from the modal rating did not impact helpfulness judgements beyond the effect of absolute deviation from the mean rating. This suggests participants did not find the most helpful reviews to be those that best represented a typical consumer's experience. Additionally, the total number of reviews did not significantly influence helpfulness judgements, which is inconsistent with Bayesian reasoning. These results provide additional support for the confirmation bias hypothesis.

Theoretical Contributions

The literature on judgements of review helpfulness is growing quickly, but it is still relatively small when considering the large role reviews play in ever-expanding online shopping. There has been limited research on how summary information influences review helpfulness. We add to this literature in several ways. First, we corroborate the results from a few papers that use field data and find the mean product rating is negatively related to helpfulness. We augment and clarify these findings by presenting causal evidence and by exploring both the processes underlying this relationship and behavioral consequences that arise from it.

We also contribute to the literature on positivity and negativity bias, which is somewhat larger than the literature on review helpfulness. As mentioned previously, this literature, in particular its overlap with work on consumer reviews and word-of-mouth, has identified many different contexts that influence whether positive or negative reviews are more helpful. Although we cannot unify all these findings, we do provide additional insight that may tie several previously observed phenomena together. The relative helpfulness of positive and negative reviews can differ depending on whether the product is hedonic or pragmatic (Sen & Lerman, 2007), experiential or material (Dai, Chan, & Mogilner, 2020), or temporally proximal or distal (Chen & Lurie, 2013). These differences all share the feature where one type of product (e.g., hedonic) leads to reviews that are more idiosyncratic to the reviewer and the other type (e.g., pragmatic) leads to reviews that are more applicable to all consumers. Our findings (and these findings from the literature) suggest positive reviews are more helpful than negative reviews for products with idiosyncratic reviews (i.e., positivity bias), while the opposite is true for products with more generally applicable reviews (i.e., negativity bias).

Lastly, this work contributes to the vast literature on confirmation biases by studying it in the context of product reviews. Studies 3 – 5 provide convincing evidence that confirmation bias is at least partially responsible for the relationship between a review's absolute deviation from the mean and its perceived helpfulness. Participants exhibited several canonical behaviors that suggest the presence of confirmation bias. This provides additional insight into how mean product ratings influence consumers' judgments and behaviors by providing process evidence that supports confirmation bias as a mediating construct.

Marketing Implications

Understanding which reviews consumers find most helpful, and which ones they are most likely to search for, can assist companies in building website designs that are maximally helpful for online shoppers. Of course, displaying reviews consumers will find more helpful is not necessarily the firm's goal. Future work may explore how the mean rating influences the reviews that drive sales. This would likely have implications for the way firms choose to present reviews of their products. Prior work has found the optimal strategy to increase sales is not simply to display all maximally positive reviews, as it leads to suspicion (Doh & Hwang, 2009).

Additionally, the differences across products in the reviews participants found helpful suggest potentially different strategies for firms depending on the product categories they produce. For example, firms that produce products with little variance in quality judgments across consumers may benefit from adopting a strategy to limit negative reviews, due to the negativity bias people express for those types of products. Conversely, firms that produce products with high variance in quality judgments may try to maximize positive reviews due to the positivity bias in those products. This could influence product development and the way firms choose to market and position their products.

Our results also suggest other potential consequences that could arise from the presence of confirmation bias in review helpfulness and search. For example, failing to search for reviews that stray from the mean may reduce a consumer's willingness to explore products with greater variance in product reviews. It could also reduce one's ability to accurately predict their utility if they fail to acquire maximally informative information, which could reduce post-purchase satisfaction and likelihood of repeat purchases. Confirmation bias could also cause consumers to terminate search too early. If consumers first seek out reviews close to the mean, which lead to greater belief updating, they may decide whether to purchase without continuing to search for reviews that stray from the mean. Future research could address some of these potential consequences of confirmation bias in review helpfulness judgments.

Limitations and Future Research

We believe this work suggests new avenues for future research on how consumers use product reviews. One important factor, outside the scope of this paper, to explore is consumers' goals or motives for reading reviews, and whether this differs based on a review's absolute deviation from the mean. Consumer A may have a goal of gathering as much information as possible before making a purchase decision. In this case, a review will be most helpful when it adds a maximal amount of new information about the product. Consumer B may have already decided whether they are likely to purchase the product or not. In this case, the consumer will be more likely to gather and/or attend to information that confirms this decision (Nickerson, 1998; Fischer, 2011). These two consumers will likely have very different criteria for determining whether a review is helpful. Future research could explore how goal setting moderates the effect of absolute deviation from the mean on review helpfulness and search.

Additionally, future work could examine if the pattern of confirmation bias we observe in helpfulness judgments increases the stickiness of product perceptions. Depending on the website design, early votes may push reviews near the mean to the top of the page. This could amplify the confirmation bias in search that we observed and cause helpfulness voting shortly after the product is listed to have lasting impacts on future sales. This would have important implications for the way companies promote products after launch and how easily firms can influence consumers' early product perceptions.

There are important limitations of our studies that are worth highlighting. First, our studies use stylized stimuli. During online shopping outside these experiments, consumers see a wide array of mean ratings, distributions, and written text. While the controlled experimental paradigms are useful to study causality and underlying cognitive processes, there is always a possibility that the specific stimuli contributed to the patterns we observe. We worked very hard to reduce the role of any one idiosyncratic stimulus by sampling several products, but it is still an infinitesimal set compared to the set of products consumers buy. Future research could examine field data across a wide array of products to explore if there are certain categories with larger or smaller effect sizes. Field experiments that randomize certain aspects of how summary information and reviews are presented could provide further insight into the robustness of the effects we observe in our studies.

Another limitation is this research does not address how the effect of absolute deviation on helpfulness and search links back to purchase behavior. In theory, a review that is more helpful should have a greater impact on purchase decisions. For example, a four-star review may increase purchase intentions significantly more when the mean is four (i.e., when the review is most helpful) than when the mean is two. We ran one study to test this hypothesis, but the results did not support it. Details of that study are in Appendix 1.A. Future research should further examine whether and/or how review helpfulness influences purchase behavior.

Conclusion

Reviews that were close to a product’s mean rating were seen as more helpful, produced greater belief changes, and were more likely to be searched than reviews far from the mean. We provide evidence that those relationships are due to confirmation bias. These results augment our understanding of how consumers use product reviews and have important implications for both marketers and researchers.

Chapter 2 – Beyond Unidimensional Sentiment Analysis: The Effect of Valence and Arousal on Evoked Emotion and Product Evaluations

Abstract

Online shopping and advances in text analysis have made researchers and marketers quite interested in the emotional content of product reviews. The vast majority of work in this space focuses solely on the valence of the review's author. We present evidence that arousal, in conjunction with valence, can significantly influence consumers' emotions, judgments, and choices. We also highlighted the need to consider the emotion experienced by the reader, as those feelings are distinct from the writer's and can significantly impact consumer behavior. We also demonstrate why it is important to measure positive and negative emotion on separate, unipolar scales, as opposed to a single, bipolar scale. These findings contribute to existing research on emotion and product reviews, and they allow researchers and marketers to gain a better understanding of how reviews impact consumer behavior.

More than 90% of consumers in the United States have used online reviews to help them make a purchase (Kaemingk, 2020). The majority of consumers trust these online reviews at least as much as they trust personal (Statista, 2021) and expert (Galante, 2018) recommendations. Consumers consider these reviews to a be a key indicator of product quality (de Langhe et al., 2016). Importantly, consumers consider the content or text of a review to be a highly influential feature of online reviews, above and beyond star ratings and total number of reviews (Podium, 2017). One key feature of this content is the emotion expressed by the person writing the review. Marketers and firms have grown wise to the fact that this emotion affects their sales, and companies now offer services focused on managing sentiment in reviews (e.g., https://www.revuze.it/). However, much remains unknown about how consumers react to this emotion.

Emotions are often depicted as an unreliable guide that leads people astray from the better judgment offered by their rational or cognitive faculties. Take, for example, a Harvard Business Review article titled: “Don’t Let Emotions Screw Up Your Decisions” (Gino, 2015). This perspective suggests consumers might discount the information presented in a highly emotional review, considering it to be less diagnostic of product quality than a more straightforward assessment. On the other hand, emotional responses are often adaptive and provide information that is useful for guiding behavior. One example is when people avoid a risky option because of the negative emotion associated with it (Bechara et al., 1997; Loewenstein et. al., 2001; Lerner et al., 2015). This perspective suggests consumers might interpret a highly emotional negative review as indicator that a product is truly low quality.

Considering the importance of emotions and product reviews in consumer behavior, there is active research examining emotions in reviews. Researchers and marketers will often use sentiment analyses to measure how positive or negative a review is. However, these analyses often use a single, unidimensional scale which yields an incomplete representation of emotion.

There is a long, robust stream of research supporting the importance measuring arousal, in addition to valence, when characterizing emotions (Barrett & Russell, 1999). This presents an opportunity to build on existing research by studying the effect of arousal in product reviews. Additionally, most research on emotion in product reviews, as well as most methods of sentiment analysis, focuses exclusively on the emotions expressed in the text, which are the emotions of the review's author. However, this does not capture the emotions experienced by prospective consumers who use those reviews to make purchase decisions. We present evidence that evoked emotion may predict consumer judgments and behaviors better than expressed emotion. While the emotions of the author and reader are likely correlated, they are distinct constructs, and we find they both have significant explanatory power in predicting behavior.

Another issue with simply measuring the valence of text is that it often uses a single, bipolar scale that ranges from negative to positive. However, this method has an important drawback which is the meaning of the midpoint of the scale is ambiguous. Responses at the midpoint could reflect reviews that are very bland or reviews that have high degrees of both positive and negative emotions. Prior literature has found evidence for positive and negative emotion acting as distinct constructs as opposed to opposite ends of the same scale (Watson, Clark, & Tellegen, 1988). We present evidence that shows it is important to measure positive and negative emotion separately, as certain patterns of behavior can only be explained with separate scales.

Lastly, while this paper supports a dimensional account of emotions, there is literature that posits emotions should be considered as discrete categories rather than being modeled as having two (or more) separate dimensions. This could be another limitation of measuring valence only, as there are many positive and negative emotional states that differ from each other but could have similar valence ratings (Bradley & Lang, 1999). While emotions may be best classified as categories, the valence-arousal framework is a robust and parsimonious model to capture differences between these various emotional states. It is important to note that the valence-arousal framework is meant to be a descriptive model of those discrete emotions. Thus, the dimensional view of emotion compliments, rather than contradicts, the categorical view of emotions (Russell, 1980). We present evidence for why it is important to consider both.

Across six studies, we find consumers use emotion expressed in reviews as a signal of product quality. Both valence and arousal influence judgments of product quality, willingness to pay (WTP), and product choice. We also find that evoked emotion (i.e., emotion experienced by participants in response to reading reviews) mediates these relationships. Furthermore, we find patterns of quality judgments that can be predicted when using unipolar scales for valence, but not when using bipolar scales. Lastly, we show how manipulating discrete emotional states can affect valence, arousal, and consumer behavior. These findings make key contributions to existing literature on emotion in product reviews by demonstrating the importance of accounting for arousal and evoked emotion.

Theoretical Background

Valence and Arousal in Consumer Reviews

There is a rich body of research on the importance of both valence and arousal in characterizing emotions. Russell (1980) posited the valence-arousal framework by modeling emotions as varying from misery to pleasure and, independently, from arousal to sleepiness. Along similar lines, Watson and Tellegen (1985) defined arousal implicitly via low and high positive or negative affect, which is a 45° rotation of the valence-arousal axes. In a later model, Barrett and Russell (1999b) theorized a two-dimensional model using pleasantness and activation as independent dimensions. A review of several dimensional models of emotion reveals that, despite slight differences, these models converge to suggest that both valence and arousal are fundamental to emotion (PS & Mahalakshmi, 2017; see Table 2.1). Additionally, prior research has found the valence and arousal of emotions primarily affect different parts of the brain (Colibazzi et. al., 2010). While some models include more dimensions, most variance in measurements of emotion can be accounted for along these two dimensions (Russell, 1980). Therefore, we will focus on valence and arousal in this paper, but note the valence-arousal framework is not the only dimensional model of emotion (for a review, see PS & Mahalakshmi, 2017).

Table 2.1: Dimensional Models of Emotion

Authors	Dimensions
Russell	Valence (pleasure) and arousal (activation)
Mehrabian	Pleasure, arousal, and dominance
Watson and Tellegen	Positive activation and negative activation
Lee et al.	Negative and non-negative emotion
Kleinsmith et al.	Valence, arousal, potency, and avoidance
Vogt et al.	Positive-active, negative-active, positive-passive, negative-passive
Khan et al.	Positive, neutral, and negative emotion
Hasan et al.	Happy-active, happy-inactive, unhappy-active, unhappy-inactive

NOTE – Adapted from PS & Mahalakshmi (2017)

The prevalence of online product reviews and new computational tools to analyze text have led to novel insights regarding how valence and/or arousal may affect consumer behavior. However, they are often examined separately. Focusing first on valence, positively-valenced consumer reviews have been shown to lead to a range of better outcomes for firms (Chintagunta et al., 2010; Kronrod & Danziger, 2013; Ludwig et al., 2013; Rocklage & Fazio, 2020). For example, Chintagunta and colleagues (2010) found the valence of user reviews for movies was a stronger predictor of box office sales than volume of reviews. In the context of word-of-mouth behavior, Ludwig and colleagues (2013) found that positive affective content predicted increases in conversion rates for Amazon book sales a week later. Consistent with these findings, a recent meta-analysis of 26 studies, using a variety of large real-world datasets, found valence in online reviews to be a strong predictor of sales elasticities (Floyd et al., 2014).

Other studies have focused exclusively on the arousal level in the text of reviews, (Yin, Bond, & Zhang, 2017; Rocklage & Fazio, 2020). One study examined the relationship between arousal and review helpfulness in reviews from Apple’s App Store (Yin, Bond, & Zhang, 2017).

Yin and colleagues (2017) demonstrated a non-linear relationship between arousal and review helpfulness. Helpfulness was relatively low for reviews with low or high levels of arousal, but high for reviews with a moderate level of arousal. However, that research does not examine how or whether this relationship is affected by valence. More recently, Rocklage and Fazio (2020) found that emotionality is positively associated with review helpfulness, but their investigation was limited to positive reviews.

While many papers have examined the role of emotion in consumer reviews, relatively few have manipulated or measured valence and arousal together. For instance, certain studies have only examined either positive reviews (Rocklage & Fazio, 2020) or negative reviews (Kim & Gupta, 2012). Other studies do not distinguish between valence and arousal, instead treating emotion as a unitary construct ranging from negative to positive (Chen & Lurie, 2013; Ludwig et al., 2013; Rocklage & Fazio, 2020). There is additional work that only examines valence, with no variation in arousal (Sen & Lerman, 2007; Schindler & Bickart, 2012). Without incorporating both dimensions, this research risks missing important insights into how consumers respond to emotion in product reviews.

Notably, high-arousal reviews are not simply more negative or more positive than their lower-arousal counterparts (Reisenzein, 1994). Arousal is defined as the level of activation that an emotion induces, and ranging from calm, or low activation, to excited, or high activation (Bestelmeyer, Kotz, & Belin, 2017). Intensity has been proposed as a third dimension to the valence-arousal model of emotions (Russell & Barrett, 1999). For example, one can feel extremely calm, which is an emotional state that has low arousal but high intensity. Thus, valence and arousal describe an emotional state relative to other states, while intensity describes the absolute level of that emotional state.

Evoked Emotion

While there is a large stream of literature on the emotions expressed in text, there is comparatively less work on how the emotions in text translate to emotions in readers (Yang, Lin, & Chen, 2009). We propose the emotion in the review is important because of how it makes the consumer feel when reading the review. While they are surely correlated, the emotion expressed by the writer and the emotion evoked in the reader are not synonymous. We focus on evoked emotion since it is considered a primary pathway by which emotional stimuli influence relevant judgments and decisions (Ajzen, 1991). We find evoked emotion mediates relationships between the emotion expressed by the writer and participants' product evaluations.

The idea that emotions contain information relevant to judgments and decisions has been widely studied (see Lerner et al., 2015 for a review). A subset of this work examines the relationship between affect and risk perceptions (Johnson and Tversky, 1983; Finucane et al., 2000; Han et al., 2000; Lerner & Keltner, 2001; Loewenstein et al., 2001; Lerner et al., 2003; Slovic et al., 2004). While purchasing a product online is not risky in the same way that base-jumping is risky, there is risk inherent in judging whether a product will be of high or low quality and deciding to purchase that product, as the consumer could purchase a product they do not like. Indeed, classic theoretical frameworks in marketing frame purchasing decisions as risky choices (Bauer, 1960; Taylor, 1974).

Researchers have proposed several, largely convergent, theoretical frameworks that describe the relationship between affect and risk perception (Lerner & Keltner, 2001; Loewenstein et al., 2001; Slovic et al., 2004). The core idea that emerges is the emotions people feel influence their judgments. For instance, the Risk-as-Feelings hypothesis, proposed by Loewenstein and colleagues (2001), proposes that emotions serve as an input to assessments of risk. Directionally speaking, this framework proposes that negative emotions increase the perception of risks, whereas positive emotions decrease the perception of risks (Johnson & Tversky, 1983; Loewenstein et al., 2001).

Research on consumer reviews has largely not examined the role that emotion experienced by the person reading the review plays in their judgments of and decisions. Although prior research has found that evoked emotion is a strong predictor of which articles are shared online (Berger & Milkman, 2016), this mechanism has received little attention in the context of consumer reviews. This omission is notable given the work outlined above that posits a primary pathway by which a stimulus influences consumer perceptions is via the effect it has on their emotions (Ajzen, 1991).

Independence of Positive and Negative Emotion

There is a mixed stream of research on whether positive and negative emotion are independent psychological constructs or if they are simply opposite ends of the same construct. One of the most commonly used ways to measure emotion in psychology is the PANAS scale, which measures positive and negative affect separately (Watson, Clark, & Tellegen, 1988). Watson, Clark, and Tellegen (1988) ran a factor analysis on a wide array of emotional states and found evidence for a two-factor solution corresponding to positive and negative emotion. Positive and negative emotions also differ distinctly on how they persist over time (Diener & Evans, 1984; Waugh et. al., 2018). They also seem to differentially vary across individuals, gender, and cultures, which suggests a level of independence (Larsen & Ketelaar, 1991; Bagozzi, Wong, & Yi, 2021).

However, other research has cast doubt on the independence of these factors and instead treat positive and negative emotions as opposite ends of a bipolar scale. Barrett and Russell (1998) used confirmatory factor analysis to test for bipolarity and found emotions were best described by two bipolar factors (pleasantness and activation). Other work has found the degree to which positive and negative emotions are independent is context-dependent (Dejonckheere et. al., 2021). Positive and negative emotion tend to become more bipolar (i.e., dependent) as the personal relevance of the emotional stimulus increases. Furthermore, the degree of correlation between positive and negative emotion can depend heavily on the chosen measurement scale (Egloff, 1998).

A potential reconciliation that has relevance to our research is that emotional valence often aligns with a bipolar scale because positive and negative emotional states often do not co-occur (Diener, 1999). However, nothing prevents them from co-occurring, and they regularly do in the context of consumer reviews, as many reviews contain both positive and negative information (Lu, Qiu, & Wang, 2021). We explore cases like this in Study 2 and demonstrate why we believe positive and negative emotion should be considered separately, as doing so provided better predictions of participants' judgments.

Categorical Models of Emotion and Basic Emotions

Part of why positive and negative emotion are distinct is because they describe discrete emotional states than can be experienced simultaneously. This has led some research to eliminate the concepts of valence and arousal altogether, and instead account for emotions as simply a set of categories (for a review, see PS & Mahalakshmi, 2017). Effective categorization is an adaptive way for consumers to organize concepts in memory (Anderson, 1991). Categorization is also an automatic process that can occur within a few seconds of seeing a stimulus (Ashby & Maddox, 2005). Furthermore, different emotional states have been shown to have distinct signatures of neural activation (Saarimäki et. al., 2015), supporting a categorical view of emotions.

To describe emotions in terms of discrete states, an obvious question is what those states should be, as there is an enormous number of emotions people experience. One theory is there are "basic emotions" that act as building blocks for all other emotional states (Solomon, 2002). These basic emotions are generally considered to be innate, universal, and adaptive (Kowalska & Wróbel, 2020). Different theories have posited the existence of a different number of basic emotions, the most common of which involves six emotional states: happiness, sadness, disgust, fear, surprise, and anger (Ekman, 1992; Saarimäki et. al., 2015). Another commonly used model characterizes emotions using eight basic states: joy, anticipation, surprise, trust, sadness, fear, disgust, and anger (Plutchik, 1962). While many theories of basic emotions have been put forth, they tend to include states similar to these (see Table 2.2).

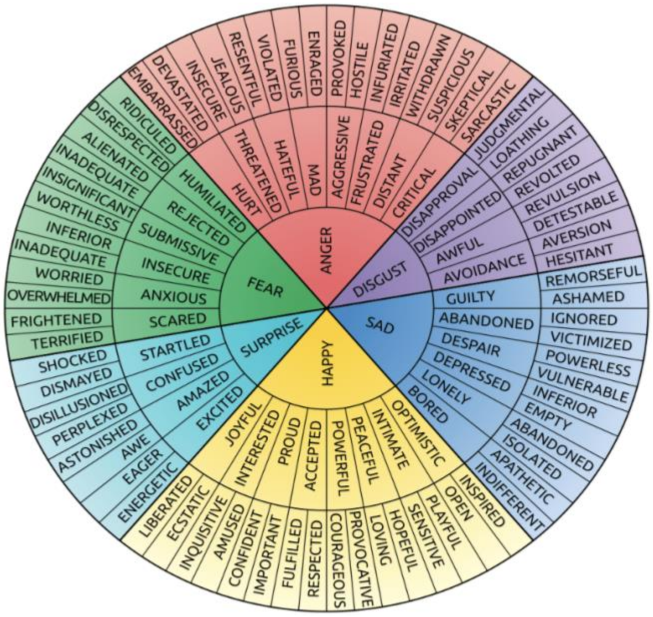
Table 2.2: Categorical Models of Emotion

Authors	**Basic Emotions**
Eckman	Anger, disgust, fear, happiness, sadness, surprise
Plutchik	Anger, disgust, fear, joy, sadness, surprise, anticipation, trust
Alm et al.	Anger, Disgust, Fear, Happiness, Sadness, Positive Surprise, Negative Surprise
Strapparava et al.	Anger, Disgust, Fear, Joy, Sadness, Surprise
Gill et al.	Anger, Fear, Surprise, Joy, Anticipation, Acceptance, Sadness, Disgust
Balahur et al.	Anger, Disgust, Fear, Guilt, Joy, Sadness, Shame
Balabantary et al.	Anger, Disgust, Fear, Happiness, Sadness, Surprise
Roberts et al.	Anger, Disgust, Fear, Joy, Sadness, Surprise, Love
Agrawal et al.	Anger, Disgust, Fear, Happiness, Sadness, Surprise
Sykora et al.	Anger, Disgust, Happiness, Sadness, Shame, Surprise, Confusion, Fear
Wang et al.	Anger, Disgust, Fear, Guilt, Joy, Sadness, Shame
Suttles et al.	Anger, Disgust, Fear, Happiness, Surprise, Trust, Anticipation, Sadness
Calvo et al.	Anger, Disgust, Fear, Joy, Sadness
Sreeja P.S et al.	Anger, Courage, Fear, Hate, Joy, Love, Peace, Sad, Surprise

NOTE – Adapted from PS & Mahalakshmi (2017)

Some empirical evidence supports the notion of basic emotions. Research in favor of basic emotions often posits that humans, universally, are born with distinct neural substrates representing a small set of basic emotions. For example, research in neuroscience has been able to map different patterns of brain activity to Plutchik’s eight basic emotions (Saarimäki et. al., 2015). Additionally, research into the taxonomy of emotions found evidence for nine basic-level categories in a hierarchical categorization task (Fehr & Russell, 1984). Basic level categories are psychologically privileged in a variety of cognitive tasks, relative to more general superordinate categories or more specific subordinate categories (Medin, Ross, & Markman, 2005). A common representation of this concept is via an emotion wheel, where basic emotions comprise the inner part of the wheel and the outer parts of the wheel are comprised of more specific states within those basic emotions (see Figure 2.1).

Figure 2.1: Emotional Wheel



Overview of Current Research

Existing research suggests emotion expressed in consumer reviews is likely to influence outcomes that are directly relevant to both consumers and firms (Chevalier & Mayzlin, 2006; Chintagunta et al., 2010; Ludwig et al., 2013; Yin et al., 2017; Rocklage & Fazio, 2020). Consistent with this literature, we expect the valence of a product review will influence product evaluations. Importantly, we propose that the valence of the emotions experienced by the reader will significantly predict quality judgments and product choice, even after controlling for the valence expressed by the writer.

H1: Positive (vs. negative) evoked emotion from a review will correspond to higher judgments of product quality, even when controlling for the valence expressed in the review.

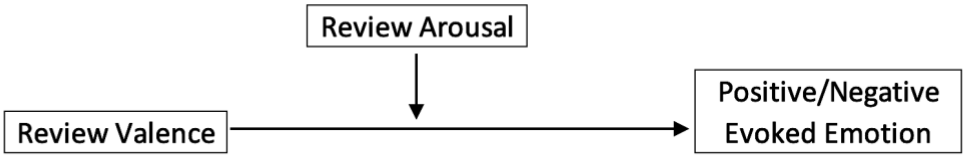
In addition to expecting evoked emotions to have explanatory power, we expect positive and negative emotion to be separate constructs. Thus, we investigate mixed reviews as a context where measuring positive and negative emotion separately can account for patterns of responses that cannot be accounted for if positive and negative emotion are measured on a single bipolar scale. We hypothesize that mixed reviews which are fairly bland affect consumers differently than mixed reviews that are emotional. Specifically, given there are individual differences in response to emotions (Larsen & Diener, 1987; Kuppens & Tong, 2010), we predict there will be greater heterogeneity in consumer responses to reviews with mixed emotions than to reviews with little-to-no emotion. We will test this by evaluating the variances of participant responses to bland and emotional reviews.

H2: For mixed reviews, the variance in quality judgments will be larger for emotional reviews than for bland reviews.

Furthermore, we expect the valence and arousal expressed in the review to jointly affect how positively or negatively a consumer feels after reading a review (Kwak, Kim, & Hirt, 2011). This is consistent with the valence-arousal framework and empirical evidence. Building on the aforementioned work on the neural signatures of valence and arousal, some work has found valence and arousal are integrated sequentially to produce emotional responses, with valence being processed first (Gianotti et. al., 2008; Colibazzi et. al., 2010). Similarly, we expect valence and arousal to jointly influence consumer emotions. Specifically, we expect higher levels of arousal to amplify the effect of valence.

H3: Negative reviews with higher (vs. lower) levels of arousal will lead to higher levels of negative evoked emotion. Positive reviews with higher (vs. lower) levels of arousal will lead to higher levels of positive evoked emotion.

Figure 2.2: Theoretical Model for Evoked Emotion



We also propose that the effects of review valence on product evaluations will be moderated by the review's level of arousal. While existing literature consistently suggests arousal in reviews is likely to influence consumer reactions, the direction of that effect remains uncertain. One possibility is that emotion would lead consumers to discount the credibility of the review writer (Petty, 2020; Karduni et al., 2021) and consider a review with higher arousal to be less diagnostic. This would lead consumers to associate products with higher (vs. lower) arousal negative reviews to be higher quality and to associate products with higher (vs. lower) arousal positive reviews to be lower quality (e.g., Kim & Gupta, 2012). However, we instead base our predictions on the literature indicating that valence tends to directionally influence outcomes relevant to firms. For example, positively valenced reviews predict better box office sales (Chintagunta et al., 2010). Correspondingly, we propose that consumers will react more strongly to reviews with higher arousal, which will translate into differences in purchase behavior.

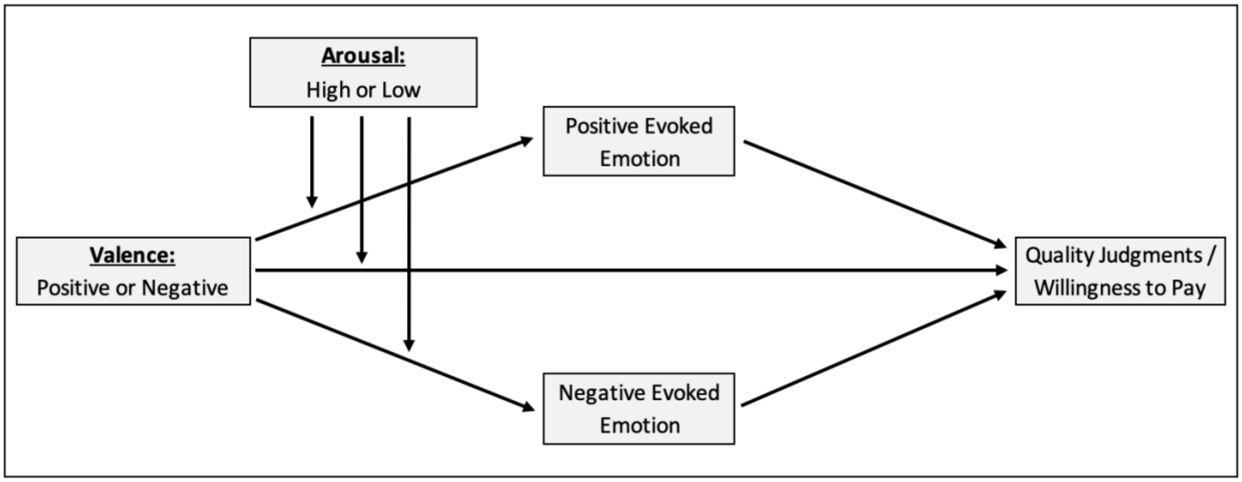
> **H4:** For positive reviews, products associated with high- (vs. low-) arousal reviews will be judged as higher quality, command a higher WTP, and receive a larger choice share. For negative reviews, arousal will have the opposite effect on perceived quality, WTP, and choice.

Our final prediction is that positive and negative evoked emotion will mediate the relationships between valence of the review and product quality judgments. This is because evoked emotion is a primary mechanism through which emotional stimuli impact behavior (Ajzen, 1991; Kwak, Kim, & Hirt, 2011). However, we expect the indirect paths to be amplified by higher levels of arousal. Thus, we propose a moderated mediation model (see Figure 2.3).

> **H5a:** Evoked positive and negative emotion will mediate the relationship between review valence and quality judgments/WTP.

> **H5b:** Review arousal will moderate the indirect paths such that the magnitude of the indirect effects is greater for products with higher- (vs. lower-) arousal reviews.

Figure 2.3: Theoretical Model for Perceived Quality and WTP



Overview of Studies

Six studies demonstrate the advantages of accounting for the level of arousal in reviews, consumers' evoked emotions, and the independence of positive and negative emotion. Study 1 uses a broad set of product reviews from Amazon and measures participants perceived quality, as well as ratings of the positive and negative emotion expressed by the writer and felt by the reader. We find evoked emotion significantly predicts quality, more so than expressed emotion. Study 2 uses a set of mixed reviews that experimentally manipulates how much positive and negative emotion is expressed in the review while randomizing the informational content. We find greater heterogeneity in quality judgments for the emotional reviews, relative to the bland reviews. Also, the emotional and bland reviews are indistinguishable on a bipolar scale but significantly different on two unipolar scales. Studies 3, 4A, and 4B manipulate the valence and arousal of reviews and their effect on quality judgments, WTP, and product choice. Higher levels of arousal tend to amplify the effect of valence on those dependent measures. Lastly, Study 5 manipulates the basic emotion expressed in reviews and measures how it affects arousal, evoked emotion, quality judgments, and WTP. We again find arousal amplifies the effect of valence. Taken together, these studies significantly enhance the fields understanding of how valence and arousal influence consumers' emotions and behaviors.

Studies 2, 4A, 4B, and 5 were preregistered. Those preregistrations and the full text for all studies are available at:

https://osf.io/wrbhf/?view_only=8a4603b1bb4548acbb0ff373d75576be

Table 2.3: Summary of Study Results

Study	Hypothesis	Result
S1	H1: Evoked emotion predicts quality judgments.	H1 supported. Positive and negative evoked emotion predicted quality judgments better than expressed emotion.
S2	H2: Greater heterogeneity in quality judgments of mixed-emotion reviews than of bland reviews.	H2 supported. The variance in quality judgments was higher for emotional reviews, which could only be predicted using unipolar scales for valence.
S3	H3: Arousal will amplify the effect of valence on quality judgments/WTP. H4: Positive and negative evoked emotion will mediate the effect of valence on quality judgments/WTP, which will be moderated by arousal.	H3 supported. For positive reviews, quality judgments, WTP, and choice shares increased with arousal. For negative reviews, the opposite was true. H4 supported. Positive and negative evoked emotion were significant mediators and arousal was a significant moderator. Note: the support was weak for WTP.
S4A	H3 and H4	H3 and H4 supported.
S4B	H3 and H4	H3 and H4 supported for positive reviews only. Arousal had no effect for negative reviews.
S5	H3	H3 supported. Different basic emotions were judged to have different levels of arousal. Those arousal ratings correlated positively with quality judgments/WTP for positive reviews but were negatively correlated for negative reviews.

Study 1: The Importance of Evoked Emotion

Study 1 used real Amazon reviews to further examine whether and how emotion in reviews is associated with quality judgments. To ensure our results are not idiosyncratic to the particular products or reviews that we use, we conducted a study using reviews from a very broad set of product categories. In addition to asking participants to rate reviews on several dimensions related to emotion, we asked participants to rate reviews on other dimensions (e.g., helpfulness) that have previously been identified as correlating with emotion and/or quality judgments.

Method

Participants. We aimed to recruit 600 participants online through the Cloud Research platform. 621 participants completed the entire study (M_{age} = 40, 55% female). Due to a technical problem, 14 participants had incomplete data for one of our control measures (the PANAS scale). We excluded those participants in our primary analyses, leaving 607 valid completions. Including those participants and excluding the PANAS scale as a control yields qualitatively identical results.

Design and Procedure. Study 1 used a 24(product; between) x 2(review valence: positive vs. negative, within) mixed design. We identified 48 reviews of top-selling products on Amazon, two reviews each for four products from six product categories: appliances, home and kitchen, electronics, sports and fitness, tools and home improvement, and entertainment (see Appendix 2.C for full list of products). Specifically, we selected the top positive and top negative review (as indexed by Amazon) for each product that was less than 400 words. Each participant was asked to read a description of one product. Next, participants were asked to read the top negative and top positive review of that product (presented in randomized order on separate pages). After reading each review, participants judged the product quality and rated the review on a variety of other dimensions.

Measures. After reading the product description and the product review, participants provided a judgment of the quality of the product (1 = worst quality; 7 = highest quality), how helpful the review was (1 = not at all helpful; 5 = extremely helpful), how objective the review was (1 = not objective or unbiased at all; 5 = extremely objective and unbiased), whether they thought the person who wrote the review was a reliable source of information (1 = not all reliable; 5 = completely reliable), how informative the review was about the product (1 = no information at all; 5 = a great deal of information), how vivid or detailed the description of the reviewer's experience with the product was (1 = not at all vivid; 5 = extremely vivid), how similar the participant felt to the reviewer (1 = not at all similar; 5 = extremely similar), and how likely the participant thought it was that he or she would have a similar experience to the reviewer if the participant were to purchase the product (1 = extremely unlikely; 7 = extremely likely). Note that all unipolar scales have five points and all bipolar scales have seven. Next, participants separately rated the negative and positive emotion expressed by the reviewer (1 = none or very little negative/positive emotion; 5 = an extreme amount of negative/positive emotion) and the negative and positive emotion they felt while reading the review (1 = none or very little negative/positive emotion; 5 = extremely negative/positive. In addition, participants completed the positive and negative affect schedule (i.e., PANAS; Watson, 1988).

Results and Discussion

Regression Analyses. We first examined whether negative and/or positive evoked emotion were significantly associated with judgments of product quality, above and beyond other features that might be associated with judgments of quality. To test this, we fit two mixed-effects linear regressions[7]. In our first model (the reduced model), we regressed quality judgments on positive and negative expressed emotion and our other control variables, omitting evoked emotion. In our second model (the full model), we add positive and negative evoked emotion to that regression. Because these two models were nested, we used a Likelihood Ratio Test to determine whether accounting for evoked emotion significantly improved model fit. To account for repeated measures across all linear mixed-effects models, we included participant- and product-level random intercepts. To check the level of multicollinearity, we calculated VIF scores with our reduced and full models. All VIF scores were under five, so we do not believe multicollinearity poses a problem when interpreting these models. This also implies that evoked and expressed emotion, while correlated, tap into distinct psychological constructs.

As expected, the negative reviews were associated with lower quality judgments and the positive reviews were associated with higher quality judgments (see Table 2.4). Consistent with H1, adding negative and positive evoked emotion to our reduced model significantly improved model fit ($\chi^2(2) = 111.54$, $p < .001$). Both negative and positive evoked emotion were significant and in the expected direction. We also fit a third model that included evoked emotion but excluded expressed emotion, which provided a better fit than the reduced model with only expressed emotion. Additionally, in the full model, negative expressed emotion becomes insignificant. These results suggest evoked emotion played a greater role in

[7] We used the R packages `lme4` (Bates et al., 2015) and `lmerTest` (Kuznetsova et al., 2017) to fit our mixed-effects regressions, using the Satterthwaite method to calculate degrees of freedom.

participants' quality judgments than expressed emotion. The only other variables that were significantly associated with judgments of quality were vividness and negative incidental emotion (from PANAS scale), but these coefficients were much smaller than our primary emotion measures (see Appendix 2.B for full regression table).

Table 2.4: Study 1 Regression Results

	Expressed Emotion Only	**Evoked & Expressed Emotion**	**Evoked Emotion Only**
Predictors	Estimates	Estimates	Estimates
Intercept	2.88 (1.61)	3.15 * (1.50)	3.85 ** (1.48)
Negative Expressed Emotion	-0.24 *** (0.03)	-0.03 (0.04)	
Positive Expressed Emotion	0.47 *** (0.03)	0.26 *** (0.04)	
Negative Evoked Emotion		-0.30 *** (0.04)	-0.40 *** (0.03)
Positive Evoked Emotion		0.26 *** (0.04)	0.45 *** (0.03)
Random Effects			
τ_{00}	0.03_{id}	0.05_{id}	0.04_{id}
	0.12_{product}	0.10_{product}	0.10_{product}
N	601_{id}	601_{id}	601_{id}
	24_{product}	24_{product}	24_{product}
Observations	1202	1202	1202
Marginal R^2	0.45	0.50	0.48
Conditional R^2	0.52	0.57	0.55
AIC	3680	3569	3612

NOTE - * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Discussion. The results of Study 1 indicate that emotion expressed in consumer reviews was strongly associated with judgments of quality. In addition, these results were relatively unaffected by controlling for several other relevant variables. We found initial evidence that evoked emotion has explanatory power in predicting judgments of product quality, even when controlling for expressed emotion. The emotion felt by the participants also predicted quality judgments better than the emotion expressed by the reviewer.

Study 2: Measuring Valence of Mixed Reviews

In Study 1, and in future studies, we measure the amount of positive and negative emotion on separate, unipolar scales, as is common in psychology research on emotions (Diener & Emmons, 1984; Watson, 1988). However, methods of sentiment analysis often measure valence on a single, bipolar scale ranging from very negative to very positive. While such a scale may work well for reviews that are unambiguously positive or negative, many reviews are mixed. Consequently, the meaning of the midpoint of the bipolar scale is ambiguous, as it could reflect reviews that have virtually no emotion or highly emotional mixed reviews. In Study 2, we examine such a case and the consequences it has for consumer behavior.

Method

Participants. We recruited 998 participants online through Prolific. One participant failed a memory check for a total of 997 participants who completed the entire study (M_{age} = 41, 48% female).

Design and Procedure. Study 2 used a 2(review type: bland, emotional; within) x 2(arousal of emotional review: high, low; between) x 2(emotion measurement scale: bipolar vs. unipolar; between) mixed design. Participants saw two book reviews. Each review has two pieces of positive information and two pieces of negative information. One review contains no emotion words and the other contains two positive and two negative emotional states (either all low-arousal or all high-arousal; see Table 2.5 for examples)[8]. For each review, participants rated the perceived quality, evoked emotion, expressed emotion, and several controls (helpfulness, objectivity, reliability, and informativeness). Bipolar scales measured evoked and expressed emotion from 1 (An extreme amount of negative emotion) to 7 (An extreme amount of positive emotion). Unipolar scales measured evoked and expressed emotion from 1 (None or very little) to 7 (An extreme amount), separately for positive and negative emotion. We present results collapsed across the arousal factor because it was generally not a significant moderator. See Appendix 2.B for analyses that do not collapse across arousal.

[8] Emotional states for Studies 2, 4A, 4B, and Supplemental Study 1 were chosen based on valence and arousal ratings from the NRC Lexicon (Mohammad & Turney, 2013).

Table 2.5: Study 2 Sample Stimuli

Review Type	**Review Text**
Bland	Average book. The plot was interesting and the author did a good job developing the characters. However, the author spent too much time in the beginning setting the scene and the writing was confusing at times.
Emotional (low arousal)	Average book. I was **grateful** that every chapter ended on a suspenseful note that made me want to keep reading and was **pleased** with the author's great use of imagery. However, I was **disappointed** because the ending wasn't great and **upset** that a lot of the dialogue felt forced.
Emotional (high arousal)	Average book. I **enjoyed** how every chapter ended on a suspenseful note that made me want to keep reading and was **thrilled** with the author's great use of imagery. However, I was **irritated** that the ending wasn't great and **hated** a lot of the dialogue felt forced.

NOTE – Participants saw only one emotional review. Non-boldface information was randomly assigned to the bland or emotional review.

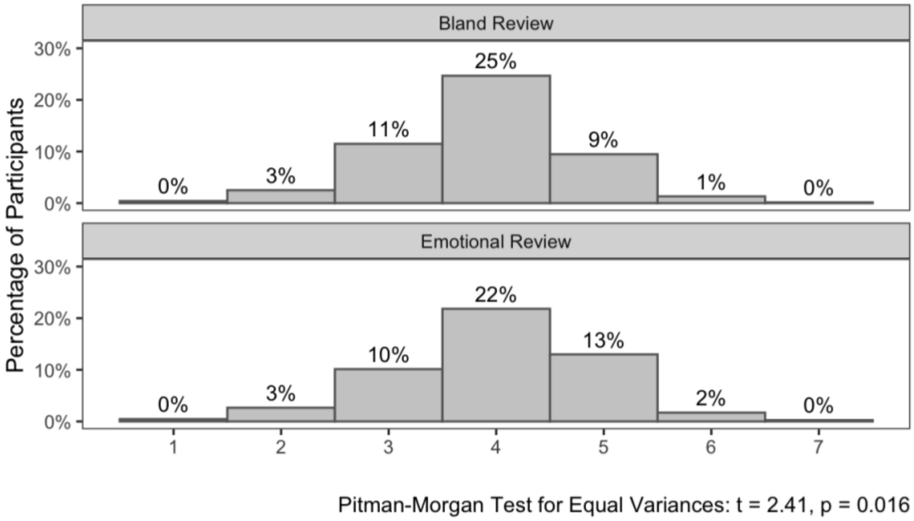
Results and Discussion

Emotion Measures. First, we ran a series of paired t-tests to compare our emotion measures between bland and emotional reviews. As expected, for participants who saw unipolar scales for valence, both positive and negative evoked emotion were greater for the emotional review than the bland review (Positive Evoked: $M_{Emotional}$ = 3.95, $SD_{Emotional}$ = 1.41, M_{Bland} = 3.57, SD_{Bland} = 1.42, t(496) = 5.63, p < .001; Negative Evoked: $M_{Emotional}$ = 3.78, $SD_{Emotional}$ = 1.36, M_{Bland} = 3.61, SD_{Bland} = 1.38, t(496) = 2.41, p = .016). The same was true for positive and negative expressed emotion (Positive Expressed: $M_{Emotional}$ = 4.43, $SD_{Emotional}$ = 1.06, M_{Bland} = 3.87, SD_{Bland} = 1.20, t(496) = 8.46, p < .001; Negative Expressed: $M_{Emotional}$ = 4.20, $SD_{Emotional}$ = 1.17, M_{Bland} = 3.80, SD_{Bland} = 1.21, t(496) = 5.49, p < .001). However, for participants who saw bipolar scales for valence, there was no difference in either evoked or expressed emotion across the two types of reviews (Bipolar Evoked: $M_{Emotional}$ = 3.87, $SD_{Emotional}$ = 0.97, M_{Bland} =

3.85, SD_{Bland} = 0.87, t(499) = 0.36, p = .716; Bipolar Expressed: $M_{Emotional}$ = 4.00, $SD_{Emotional}$ = 0.99, M_{Bland} = 3.95, SD_{Bland} = 0.92, t(499) = 0.97, p = .330)[9].

Quality Judgments. To compare the variance in product quality judgments across the two review types, we conducted a Pitman-Morgan test for equal variances[10] (see Figure 2.4). As expected, the variance of the quality judgments was greater for emotional reviews (0.94) than bland reviews (0.81; t(995) = 2.41, p = .016). While the effect is small, this suggests there is greater heterogeneity in product quality perceptions when reviews are more emotional.

Figure 2.4: Study 2 Quality Judgments



[9] The same results held in a norming study that was used to select the stimuli.

[10] As robustness checks, we ran several non-parametric tests (McCulloch, Grambsch, Bonettseier, and Levene tests) for scale differences and found the same results.

Evoked-Expressed Correlations[11]. Given that expressed and evoked emotion are likely highly correlated, we examined the correlation between the two separately for emotional reviews and bland reviews. The correlation was greater for bland reviews (0.68) than emotional reviews (0.52, z = 5.88, p < .001). This was true for both positive and negative emotion. This suggests that evoked and expressed emotion diverge more when reviews are more emotional. Also, echoing the results from Study 1, the correlations are low enough that we feel confident evoked and expressed emotion are indeed tapping into two different constructs.

Discussion. Study 2 yielded several important findings. First, we demonstrated the potential shortcomings of using a bipolar scale to measure emotional valence in reviews. While a bipolar scale showed no difference between the emotion ratings across the two review types, unipolar scales revealed the mixed reviews had both more positive and negative emotion. Second, we showed an important consequence for consumer behavior by revealing there was greater variance in quality judgments for the emotional reviews than the bland reviews. Third, we found evoked and expressed emotion tended to diverge more for more emotional reviews. Given we found evoked emotion to be a better predictor of quality judgments in Study 1, considering only the emotion expressed by the reviewer could lead to significantly worse predictions of consumer behavior if the product receives highly emotional reviews.

Arousal was not a major factor in this study, which is likely due to the high- or low-arousal emotions “cancelling out” in the context of mixed reviews. This also suggests the effect on the variance found in Study 2 cannot be accounted for by using a bipolar scale for valence

[11] Note: This analysis was not preregistered.

along with a measure of arousal, otherwise we would expect to see significant differences in variance across the high- and low-arousal conditions, which we do not (F(495, 500) = 1.00, p = .96). This suggests it is important to consider both positive and negative emotion separately, but this does not mean arousal is unimportant. Studies 3 – 5 will highlight the importance of considering arousal, in addition to valence.

Study 3: The Effect of Arousal on Purchase Behavior

Prior research suggests consumers use emotion in product reviews as a valid cue when forming judgments about a product's quality. Most sentiment analysis and research on product reviews characterize the emotion in reviews as being either positive or negative. While the valence of reviews is clearly important and can have economically significant effects on consumer behavior (Chintagunta, Gopinath, & Venkataraman, 2010), there are other important aspects of emotions that cannot be captured by measuring valence alone.

In Study 3, we manipulated both valance and arousal in reviews. We constructed a set of reviews, modeled on existing Amazon reviews, with varying levels of valence and arousal (see Table 2.6 for sample stimuli, full stimuli in Appendix 2.C). We then examined how presenting participants with reviews that differed in valence and arousal affected their quality judgments, WTP, and product choice. Finally, we tested whether evoked emotion mediates the relationship between valence and perceived quality/WTP.

Method

Participants. We recruited 700 participants online through the Cloud Research platform. Seven participants failed an attention check, leaving a total of 693 participants who completed the entire study (M_{age} = 41, 52% female).

Design and Procedure. This study implemented a 2(valence: negative vs. positive) x 2(arousal: low vs. high) within-subjects design. The low-arousal reviews were largely unemotional and focused primarily on conveying either a negative or positive evaluation of the product. High-arousal reviews used more emotional language and exclamation points to increase arousal. Participants were asked to imagine they were planning to purchase a version of two different types of products (a foldable exercise bike and a toaster oven). They were told they would read descriptions and reviews of two different products within each type (for a total of four products and reviews). Participants were told each product generally had good reviews but also some bad reviews, as that is the typical distribution of online reviews (Hu, Zhang, & Pavlou, 2009). They were asked to imagine they were reading reviews of each product and came across the ones shown in the study (see Appendix 2.C for full text of reviews).

Table 2.6: Study 3 Sample Stimuli

Arousal	**Exercise Bike Reviews (negative valence)**
Low	**Review Title**: Issues with bike seat **Review**: The assembly for this bike was relatively straightforward. The bike is fairly compact. However, when setting it up noticed that the seat is at a bit of a weird angle and is somewhat misaligned with the frame. Once I tried it out, I realized that this results in an awkward riding position and the seat also wobbles even when the knob is fully tightened. I followed the instructions very closely and reviewed them after I encountered this issue. I'm not sure the seat is constructed properly.
High	**Review Title**: Issues with bike seat. Unbelievable. Very upset! **Review**: The assembly for this bike was relatively straightforward. The bike is fairly compact. However, when setting it up noticed that the seat is at a bit of a weird angle and is somewhat misaligned with the frame. Once I tried it out, I realized that this results in an awkward riding position that I really hate. The seat also wobbles even when the knob is fully tightened, so I don't enjoy riding the bike. I followed the instructions very closely and reviewed them after I encountered this issue, but nothing changed, so that was useless. I'm not so sure the seat is constructed properly. I am so upset. I was feeling so good about doing something good for my body and now I'm just frustrated that I purchased it.

Participants first read a description of a product, then read a review of the product, then answered questions about the product and the review. Participants repeated this procedure for the other product within a given product type (i.e., a second toaster or second exercise bike). Next, participants made a forced choice between the two products within that product type. Participants then repeated this procedure for the other product type and valence. We also measured and controlled for review helpfulness, objectivity, and whether the emotions made participants disregard any information.

Measures. After reading the product description and review, participants rated the quality of the product, gave their WTP, and answered questions regarding evoked and expressed emotion. Lastly, they rated how helpful the review, how objective the review was, and whether the emotional content of the review caused them to discount the other information in the review. We included these measures as controls since they are likely to impact how the reviews affect quality judgments. We omit discussion of the control variables in the main text. While some were significant in the regressions, none of them affected our qualitative conclusions. Please see the Appendix 2.B for the full regression outputs.

Results and Discussion

Quality Judgments. First, we assessed whether valence and arousal impacted judgments of quality. We regressed quality judgments on dummy variables for the valence and arousal conditions and their interaction, as well as our control measures, a participant-level random intercept, and a product-level random intercept. We observed a significant interaction between the valence and arousal conditions ($\beta_{Interaction} = 0.66$, $t(2170) = 9.60$, $p < .001$, see Figure 2.5). For positively valenced reviews, reviews with high arousal led to higher quality judgments than those with low arousal ($M_{High} = 5.73$, $SD_{High} = 0.88$ vs. $M_{Low} = 5.40$, $SD_{Low} = 0.82$; $z = 6.97$, $p < .001$)[12]. For negatively valenced reviews, the opposite pattern emerged ($M_{High} = 2.77$, $SD_{High} = 1.14$ vs. $M_{Low} = 3.09$, $SD_{Low} = 1.10$; $z = 6.33$, $p < .001$)). Thus, arousal significantly moderated the effect of valence on quality judgments.

[12] In Studies 3 – 5, the test statistics and results in parentheses are hypothesis tests of combinations of coefficients from the full regression model, not individual z- or t-tests.

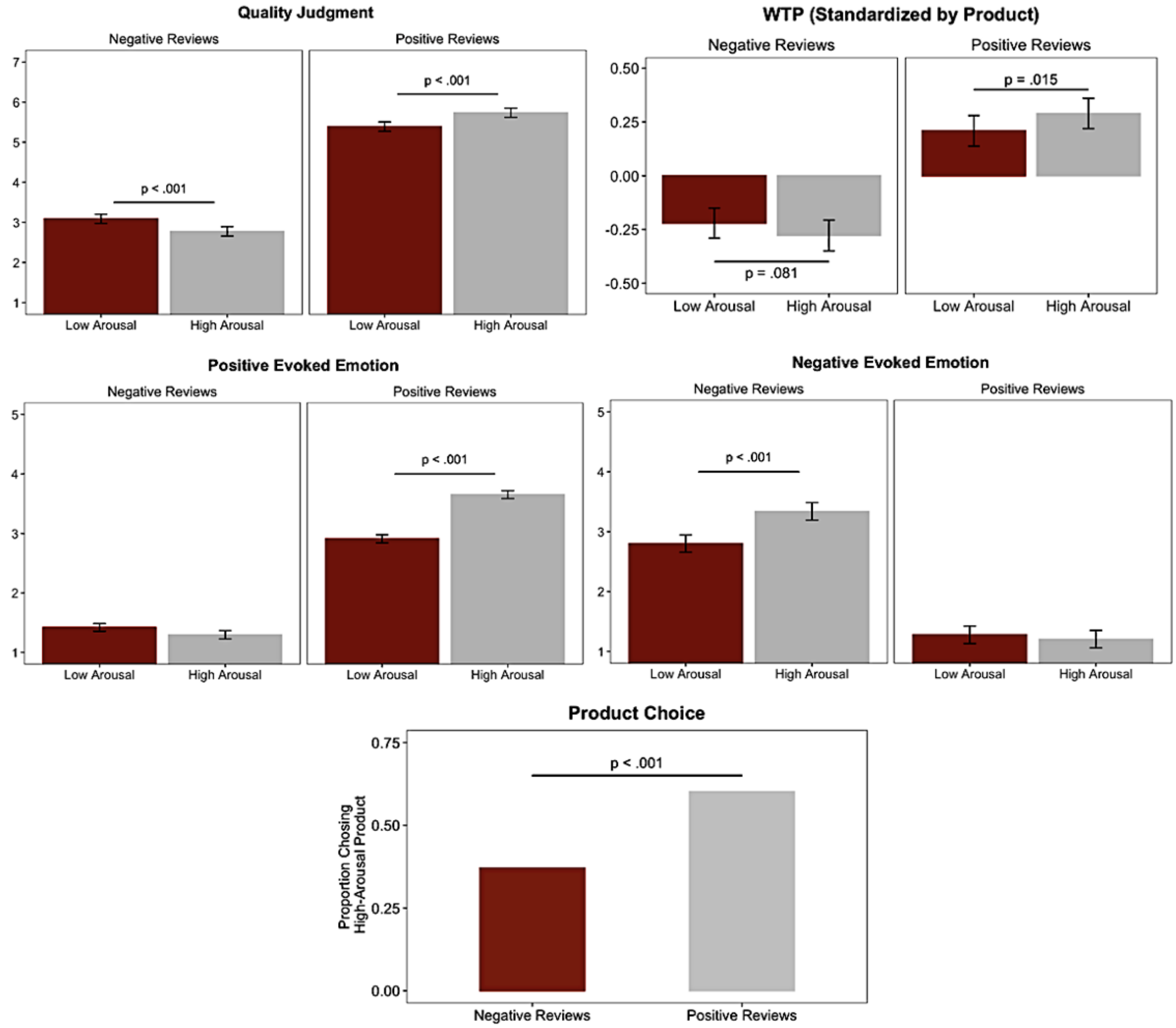
WTP. We ran the same mixed-effects regression for WTP and found similar, but weaker, results. Because the WTP scale was very different for the two products, we standardized WTP across subjects, separately for each product. We observed a significant interaction between the valence and arousal conditions ($\beta_{\text{Interaction}}$ = 0.14, t(2164) = 3.05, p = .002, see Figure 2.5). For positively valenced reviews, reviews with high arousal led to higher quality judgments than those with low arousal (M_{High} = 0.28, SD_{High} = 1.01 vs. M_{Low} = 0.19, SD_{Low} = 0.96; z = 2.44, p = .015). For negatively valenced reviews, the opposite pattern emerged (M_{High} = -0.26, SD_{High} = 0.95 vs. M_{Low} = -0.22, SD_{Low} = 0.95; z = -1.74, p = .081). Thus, arousal significantly moderated the effect of valence on WTP.

Evoked Emotion. Next, we examine the effect of valence and arousal on our measures of positive and negative evoked emotion. We ran the same mixed-effects regression, using positive and negative evoked emotion as dependent variables. We again observed a significant interaction between the valence and arousal conditions (Positive Evoked: $\beta_{\text{Interaction}}$ = 0.87, t(2092) = 14.75, p < .001, Negative Evoked: $\beta_{\text{Interaction}}$ = -0.61, t(2176) = -10.60, p < .001). For positively valenced reviews, positive evoked emotion was greater for reviews with high arousal than those with low arousal (M_{High} = 3.66, SD_{High} = 1.03 vs. M_{Low} = 2.89, SD_{Low} = 1.08; z = -17.43, p < .001). For negatively valenced reviews, negative evoked emotion was greater for reviews with high arousal than those with low arousal (M_{High} = 3.41, SD_{High} = 1.09 vs. M_{Low} = 2.82, SD_{Low} = 1.07; z = 12.91, p < .001). Thus, arousal was able to moderate the effect of valence on the evoked emotions felt by the reader.

Product Choice. Lastly, we found the same pattern for product choice. The proportion of participants who chose the high-arousal product was much higher when review valence was positive (60%) than when it was negative (37%; χ2 = 39.86, p < .001, Cohen's w = 0.24).

Moderated Mediation. To test the role of evoked emotion in explaining the effect of valence on quality judgments, we conducted a moderated mediation analysis with quality judgments as the dependent variable, review valence as the independent variable, positive and negative evoked emotion as mediators (in parallel), and review arousal moderating the both a-paths and the c-path. All paths included our control variables as well. We found negative and positive evoked emotion mediated the effect of valence on quality judgments. The indirect effects via positive and negative evoked emotion were significant at all levels of arousal (see Appendix 2.B for full mediation output). Crucially, the magnitudes of the indirect effects were greater for reviews with high arousal relative to those with low arousal (Bootstrapped 95% CI's: Index of Moderated$_{\text{via Positive Evoked}}$ = [0.33, 0.47], Index of Moderated$_{\text{via Negative Evoked}}$ = [0.16, 0.27]). This suggests a potential causal pathway via evoked emotion that is significantly moderated by the level of arousal. We ran the same analysis with WTP as the dependent variable and found positive evoked emotion to be a significant mediator at both levels of arousal, and arousal amplified the indirect effect (Bootstrapped 95% CI's: Index of Moderated$_{\text{via Positive Evoked}}$ = [0.11, 0.22]). However, negative evoked emotion was not a significant mediator (see Appendix 2.B for full output). This is likely due to the fact that the effect of arousal on WTP was only marginally significant for negative reviews.

Figure 2.5: Study 3 Results



NOTE - Error bars are 95% confidence intervals from mixed-effects regression

Discussion. In Study 3, we demonstrated that experimentally manipulating the valence and arousal of product reviews led to changes in evoked emotion, judgments of product quality, WTP, and product choice. Arousal amplified the effect of valence such that reviews with higher arousal yielded higher quality judgments, WTP, and choice shares for positively valenced reviews, while the opposite was true for negatively valenced reviews. Put differently, the difference in quality judgments, WTP, and choice shares between positive and negative reviews is larger when arousal is high than when it is low. Study 2 also highlights the importance of evoked emotion, which significantly mediated the effect of valence on quality judgments.

These findings stand in contrast to some past literature on emotion and source credibility, which has found people associate expression of emotion with lack of credibility in news reports (Karduni et al., 2021) and more generally that sources who seem emotionally biased are seen as less credible (Petty, 2020). Taken together, these results provide evidence that, rather than discounting the high-arousal reviews, participants integrated those emotions into their judgments and decisions.

Study 4A: Discrete Emotional States and Purchase Behavior

Study 3 manipulated valence and arousal in a manner we felt was externally valid and controlled for key potential confounds (e.g., helpfulness). However, there is a limitation in that, via our arousal manipulation, we may have also manipulated the intensity of the emotions in the review. Intensity has been proposed as a third dimension to the valence-arousal model of emotions (Russell & Barrett, 1999). Additionally, the valence-arousal framework is meant to classify discrete emotions, which can all take on a level intensity. In Study 4A, we manipulated valence and arousal using discrete emotional states, holding intensity of the emotions constant.

Another potential limitation of Study 3 is that the information in each review is not held completely constant. While we attempt to control for key covariates, there may be other reasons why the arousal manipulation impacted perceived quality, WTP, and choice. Study 4A addresses these limitations by holding intensity constant and randomizing the text in the review.

Method

Participants. We recruited 349 participants online through Prolific (M_{age} = 41, 48% female). For analyses involving willingness to pay (WTP), we excluded 6 outliers via Tukey's Rule (i.e., we removed observations greater than the 75th percentile plus 1.5*IQR), leaving 343 participants for those analyses.

Design and Procedure. This study used a 2(review valence: positive vs. negative, between) x 2(arousal: low vs. high, within) mixed design. Participants saw two book reviews that were either both positive or both negative. One review was high-arousal and one was low-arousal, presented in a random order. For each book, participants rated the perceived quality, their maximum WTP, positive and negative evoked emotion, and chose which book they preferred.

The way we manipulated valence and arousal in Study 4A gives us much greater internal validity by holding the vast majority of the reviews constant. First, we constructed a set of two positive reviews and two negative reviews. We also chose two positive emotions with varying arousal (high-arousal = happy, low-arousal = content) and two negative emotions with varying arousal (high-arousal = angry, low-arousal = upset). Participants saw two reviews, both with the headline "I was [emotion] with the quality of this book." For each participant that saw positive reviews, we randomly assigned which positive review went with the happy headline and which went with the content headline. For each participant that saw negative reviews, we randomly assigned which negative review went with the angry headline and which went with the upset headline. This design allowed us to isolate the effect of manipulating the discrete emotional state of the review, without potential confounds of intensity or the specific information conveyed.

Results and Discussion

Quality Judgments. First, we regressed quality judgments on dummy variables for review valence, arousal, and their interaction (clustering standard errors by participant). Again, we observed a significant interaction between the valence and arousal conditions ($\beta_{\text{Interaction}}$ = 0.30, t(348) = 3.01, p = .003, see Figure 2.6). For positively valenced reviews, reviews with high arousal led to higher quality judgments than those with low arousal (M_{High} = 5.20 , SD_{High} = 0.85 vs. M_{Low} = 5.02, SD_{Low} = 0.87; t(348) = 2.49, p = .013). For negatively valenced reviews, the opposite pattern emerged (M_{High} = 2.49, SD_{High} = 0.91 vs. M_{Low} = 2.61, SD_{Low} = 0.89; t(348) = -1.74, p = .083).

WTP. Next, we ran the same regression with WTP as the dependent variable. As with quality judgments, we found the same interaction between the valence and arousal conditions ($\beta_{\text{Interaction}}$ = 1.60, t(342) = 3.01, p = .003, see Figure 2.6). For positively valenced reviews, reviews with high arousal led to higher WTP (M_{High} = 13.35 , SD_{High} = 6.74 vs. M_{Low} = 12.53, SD_{Low} = 6.13; t(342) = 2.55, p = .011). For negatively valenced reviews, the opposite pattern emerged (M_{High} = 5.49, SD_{High} = 4.35 vs. M_{Low} = 6.28, SD_{Low} = 4.60; t(342) = -3.25, p = .001).

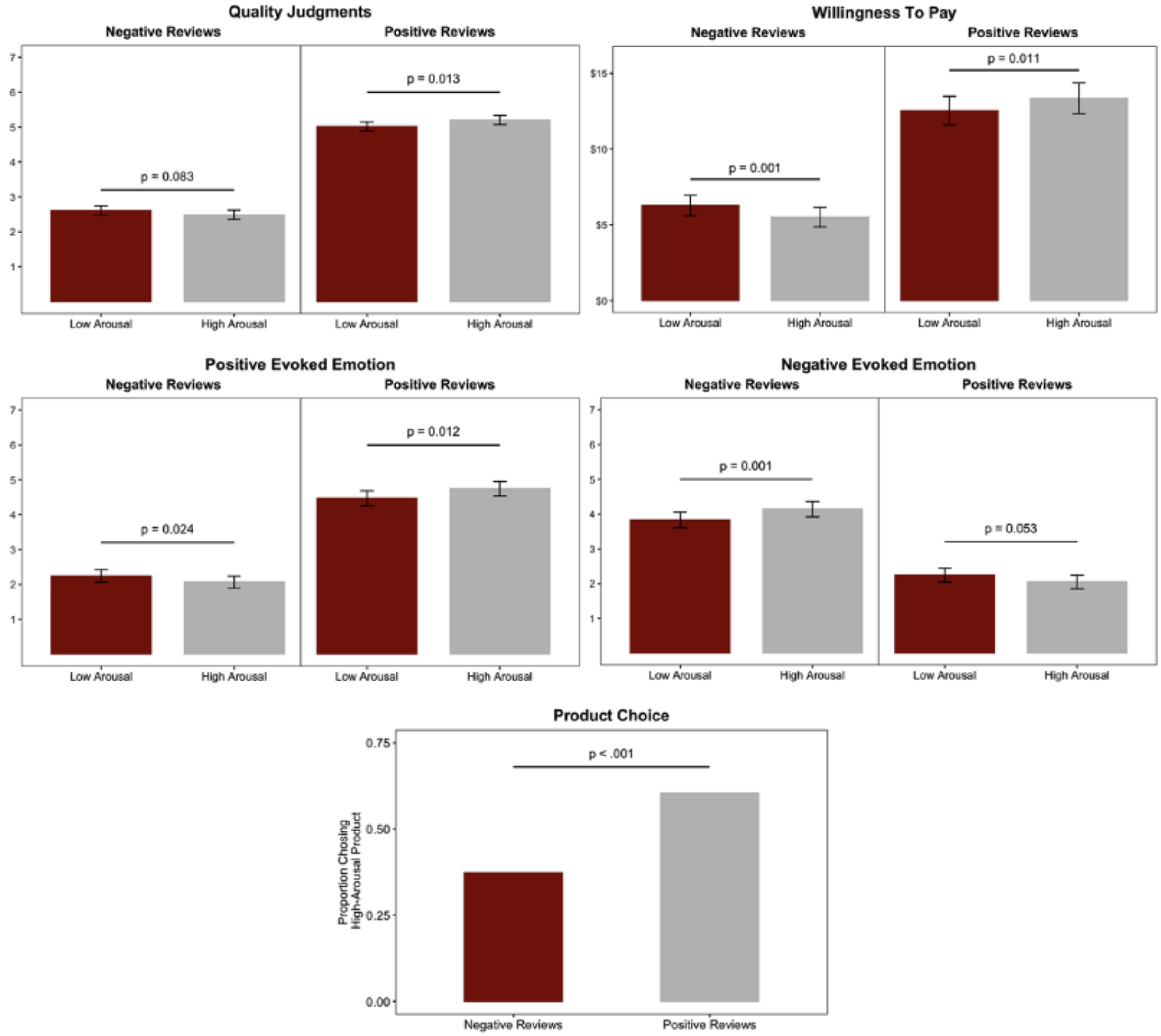
Evoked Emotion. We replicated this interaction for both positive and negative evoked emotion (Positive Evoked: $\beta_{\text{Interaction}}$ = 0.45, t(348) = 3.39, p = .001, Negative Evoked: $\beta_{\text{Interaction}}$ = -0.51, t(348) = -3.73, p < .001). For positively valenced reviews, positive evoked emotion was greater for reviews with high arousal than those with low arousal (M_{High} = 4.74, SD_{High} = 1.41 vs. M_{Low} = 4.47, SD_{Low} = 1.45; t(348) = 2.54, p = .012). For negatively valenced reviews, negative evoked emotion was greater for reviews with high arousal than those with low arousal (M_{High} = 4.15, SD_{High} = 1.50 vs. M_{Low} = 3.83, SD_{Low} = 1.53; t(348) = 3.41, p = .001).

Product Choice. Lastly, we replicated the same pattern for product choice. The proportion of participants who chose the high-arousal product was significantly higher when review valence was positive (64%) than when it was negative (39%; χ2 = 17.84, p < .001, Cohen's w = 0.23).

Moderated Mediation. With quality judgments as the dependent variable, we ran the same moderated mediation analysis as in Study 3 and replicated the results. The indirect effects via positive and negative evoked emotion were significant at all levels of arousal (see Appendix 2.B for full output), but the absolute magnitudes of the indirect effects were greater for reviews with high arousal relative to those with low arousal (Bootstrapped 95% CI's: Index of $\text{Moderated}_{\text{via Positive Evoked}}$ = [0.01, 0.19], Index of $\text{Moderated}_{\text{via Negative Evoked}}$ = [0.01, 0.11]). This, again, suggests a potential causal pathway via evoked emotion that is significantly moderated by the level of arousal. We found the same result when using WTP as the dependent variable

(Bootstrapped 95% CI's: Index of Moderated$_{\text{via Positive Evoked}}$ = [0.04, 0.56], Index of Moderated$_{\text{via Negative Evoked}}$ = [0.02, 0.46]; see Appendix 2.B for full output).

Figure 2.6: Study 4A Results



NOTE - All error bars are 95% confidence intervals (standard errors clustered by participant)

Discussion. In Study 4A, we were able to replicate all results from Study 3 with an alternative manipulation of valence and arousal that provides greater internal validity. Again, the effect of valence on various measures of purchase behavior was mediated by differences in positive and negative evoked emotion and amplified by arousal.

Study 4B: Discrete Emotional States and Purchase Behavior

Study 4A suggests the valence and arousal of the discrete emotion expressed in the headline of a review jointly influence quality judgments, WTP, and product choice. However, there are two potential limitations. First, it uses a fairly limited set of emotions. Study 4B addresses this by sampling from a set of 16 emotions. Another potential limitation in Study 4A is we did not vary the product. Thus, Study 4B uses a different product category (blenders).

Method

Participants. We recruited 800 participants online through Prolific (M_{age} = 38, 48% female).

Design and Procedure. This study used a 2(review valence: positive vs. negative, between) x 2(arousal: low vs. high, within) mixed design. Participants saw two blender reviews that were either both positive or both negative. One review was high-arousal and one was low-arousal, presented in a random order. For each blender, participants rated the perceived quality, their WTP, positive and negative evoked emotion, and chose which blender they preferred.

The manipulation of valence and arousal via changing the discrete emotional state was virtually identical to that of Study 4A. The key difference is that we used a broader set of emotions. Study 4B uses four reviews for each level of valence*arousal, yielding a total of 16 possible emotional states (see Table 2.7 for full list). Participants saw two reviews, both with the headline "I was [emotion] with/by the quality of this blender." Participants were randomly assigned one low-arousal emotion and one high-arousal emotion (both positive or both negative).

Table 2.7: Study 4B Emotional States

Valence	**Arousal**	**Emotions**
Negative	Low	disappointed, saddened, unhappy, dissatisfied
	High	angry, frustrated, furious, irritated
Positive	Low	content, pleased, satisfied, grateful
	High	happy, thrilled, excited, delighted

Results and Discussion

Quality Judgments. First, we regressed quality judgments on dummy variables for review valence, arousal, their interaction, a random intercept for the specific emotion, and a participant-level random intercept. We observed a marginally significant interaction between the valence and arousal conditions ($\beta_{\text{Interaction}}$ = 0.27, t(8.81) = 2.22, p = .054, see Appendix 2.B for full regression table). For positively valenced reviews, reviews with high arousal led to higher quality judgments than those with low arousal (M_{High} = 5.46 , SD_{High} = 0.85 vs. M_{Low} = 5.29, SD_{Low} = 0.86; z = 2.07, p = .039). However, for negative reviews, we did not observe a significant effect of arousal on perceived quality (M_{High} = 2.24, SD_{High} = 1.01 vs. M_{Low} = 2.34, SD_{Low} = 1.02; t(348) = -1.07, p = .283).

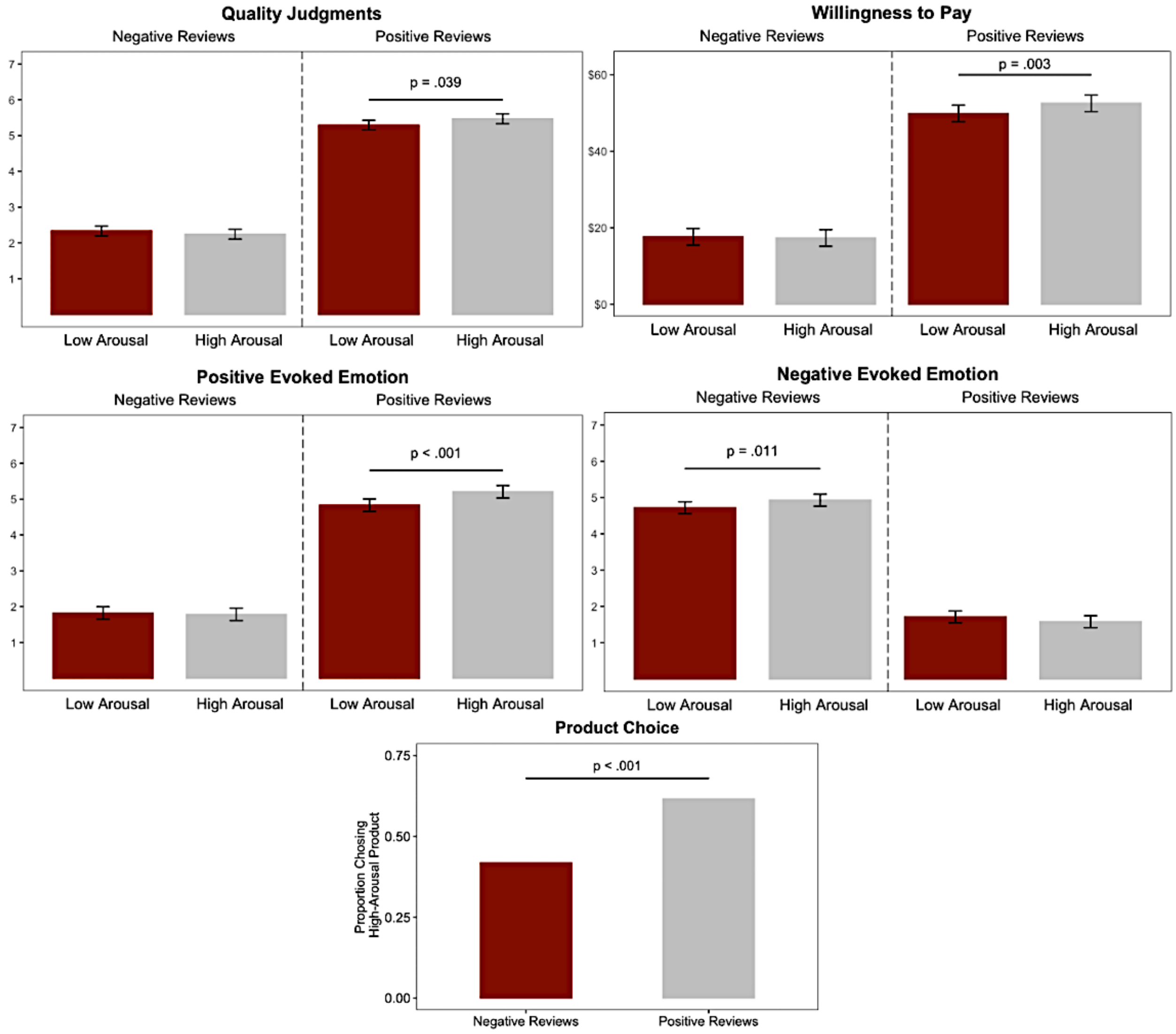
WTP. Next, we ran the same mixed-effects regression with WTP as the dependent variable. As with quality judgments, we found a marginally significant interaction between the valence and arousal conditions ($\beta_{Interaction}$ = 2.98, t(6.82) = 2.33, p = .054, see Appendix 2.B for full regression table). For positively valenced reviews, reviews with high arousal led to higher WTP (M_{High} = 52.54 , SD_{High} = 23.81 vs. M_{Low} = 49.85, SD_{Low} = 23.18; z = 2.96, p = .003). For negatively valenced reviews, there was no significant effect of arousal on WTP (M_{High} = 17.34, SD_{High} = 14.84 vs. M_{Low} = 17.64, SD_{Low} = 14.54; z = -0.33, p = .742).

Evoked Emotion. We replicated this interaction for both positive and negative evoked emotion (Positive Evoked: $\beta_{Interaction}$ = 0.64, t(10.46) = 3.41, p = .001, Negative Evoked: $\beta_{Interaction}$ = -0.71, t(10.60) = -3.29, p = .008; see Appendix 2.B for full regression tables). For positively valenced reviews, positive evoked emotion was greater for reviews with high arousal than those with low arousal (M_{High} = 5.21, SD_{High} = 1.25 vs. M_{Low} = 4.83, SD_{Low} = 1.39; z = 3.50, p < .001). For negatively valenced reviews, negative evoked emotion was greater for reviews with high arousal than those with low arousal (M_{High} = 4.93, SD_{High} = 1.68 vs. M_{Low} = 4.72, SD_{Low} = 1.70; z = 2.53, p = .011).

Product Choice. Lastly, we replicated the same pattern for product choice. The proportion of participants who chose the high-arousal product was significantly higher when review valence was positive (59%) than when it was negative (38%; χ2 = 30.50, p < .001, Cohen’s w = 0.20).

Moderated Mediation. With quality judgments as the dependent variable, we ran the same moderated mediation analysis as in Studies 3 and 4A and replicated the results. The indirect effects via positive and negative evoked emotion were significant at all levels of arousal (see Appendix 2.B for full output), but the absolute magnitudes of the indirect effects were greater for reviews with high arousal relative to those with low arousal (Bootstrapped 95% CI’s: Index of $\text{Moderated}_{\text{via Positive Evoked}}$ = [0.04, 0.14], Index of $\text{Moderated}_{\text{via Negative Evoked}}$ = [0.01, 0.07]). We found the same result when using WTP as the dependent variable (Bootstrapped 95% CI’s: Index of $\text{Moderated}_{\text{via Positive Evoked}}$ = [0.43, 1.83], Index of $\text{Moderated}_{\text{via Negative Evoked}}$ = [0.00, 0.62]; see Appendix 2.B for full output).

Figure 2.7: Study 4B Results



NOTE - All error bars are 95% confidence intervals (standard errors clustered by participant)

Discussion. In Study 4B, we were able to replicate some of the results from Studies 3 and 4A. As before, the effect of valence on various measures of purchase behavior was mediated by differences in positive and negative evoked emotion and amplified by arousal. Additionally, we generally replicated the valence-arousal interaction for our key measures. However, arousal only affected quality judgments and WTP for positive reviews, not for negative reviews. This could be due to the specific emotions we used or the different product type. There is some research suggesting arousal matters more for reviews of hedonic products (e.g., books) than pragmatic products (e.g., blenders), which could potentially explain why the pattern was stronger for books (Ren & Nickerson, 2019). However, we do still observe a similar, albeit weaker, pattern using blenders.

Study 5: The Impact of Basic Emotions on Valence, Arousal, and Product Evaluations

Studies 4A and 4B showed the power of discrete emotional states in terms of impacting consumer behavior. Study 5 investigates this further by building on theories of basic emotions. We also used an alternative manipulation that may be more externally valid than Studies 3, 4A, and 4B. The valence-arousal framework was developed to describe the relationships between basic emotional states that underly all other emotions (Russell, 1980). Thus, we expect the levels of valence and arousal present in each basic emotion to lead to similar patterns as the ones we observed in prior studies.

In Study 5, we take advantage of the publicly available NRC emotion lexicon that indicates whether a particular English word is associated with a given basic emotion (Mohammad & Turney, 2013). This lexicon utilizes eight basic emotions identified by Plutchik (1962): joy, anticipation, surprise, trust, sadness, fear, disgust, and anger. We use those eight basic emotions in Study 5.

Method

Participants. We recruited 800 participants online through Prolific. One participant failed a memory check, leaving a total of 799 participants (M_{age} = 39, 48% female). For analyses involving willingness to pay (WTP), we excluded 60 outliers via Tukey's Rule, leaving 739 participants for those analyses.

Design and Procedure. Study 5 used a 2(valence: positive vs. negative, between) x 4(basic emotion: joy, anticipation, surprise, and trust or sadness, fear , disgust, and anger; within) mixed design. In other words, participants saw all four positive emotions or all four negative emotions, in a random order. For each review participants saw, we measured perceived quality, WTP, positive and negative evoked emotion, and arousal (scale from Berger, 2011). The arousal scale has three items and asks if the review is: (i) 1 = very passive, 7 = very active, (ii) 1 = very mellow, 7 = very fired up, and (iii) 1 = very low-energy, 7 = very high energy. We also collected review helpfulness, objectivity, reliability, and informativeness as control variables.

To generate reviews for each emotion, we used the NRC Lexicon and GPT-4o to construct reviews. For each emotion, we gathered a list of all words in the lexicon associated with that emotion. We then prompted GPT-4o to construct a book review using only the words associated with that emotion, in addition to function words (e.g., and, the, etc.) and book-related words (e.g., plot, author, etc.). We did this three times for each emotion for stimulus diversity and participants were randomly assigned one of the three reviews for a given emotion (See Table 2.8 for example stimuli, the full set of reviews is in Appendix 2.C).

Table 2.8: Study 5 Sample Stimuli

Basic Emotion	Review Text
Anger	This book, despite some interesting moments, ultimately left me disappointed. The story felt chaotic and often confusing. The characters were plagued by adversity, constantly facing conflict and hardship. The narrative was filled with aggressive confrontations and antagonistic interactions. Although some parts had potential, the overall execution fell short, leading to an unsatisfying experience.
Anticipation	This was a great book filled with suspense and excitement. Characters are relatable and admirable, adding depth to the story. Although some parts feel overly ambitious, the overall experience is thrilling and enjoyable. The plot keeps you engaged with tension and anticipation, making it hard to put down. Despite some clear flaws, it's a spectacular book.

NOTE - See Appendix 2.C the full set of reviews

Results and Discussion

Arousal Ratings. First, we calculated participants' arousal ratings for each emotion as the average of the 3-item scale (average Cronbach's alpha = 0.87). Separately for positive and negative emotions, we conducted pairwise comparisons for arousal ratings between each basic emotion, all of which were significant (p's ≤ .001, see Figure 2.8). For positive emotions, surprise had the highest average arousal, followed by anticipation, joy, and then trust. For negative emotions, disgust had the highest average arousal, followed by anger, fear, and sadness.

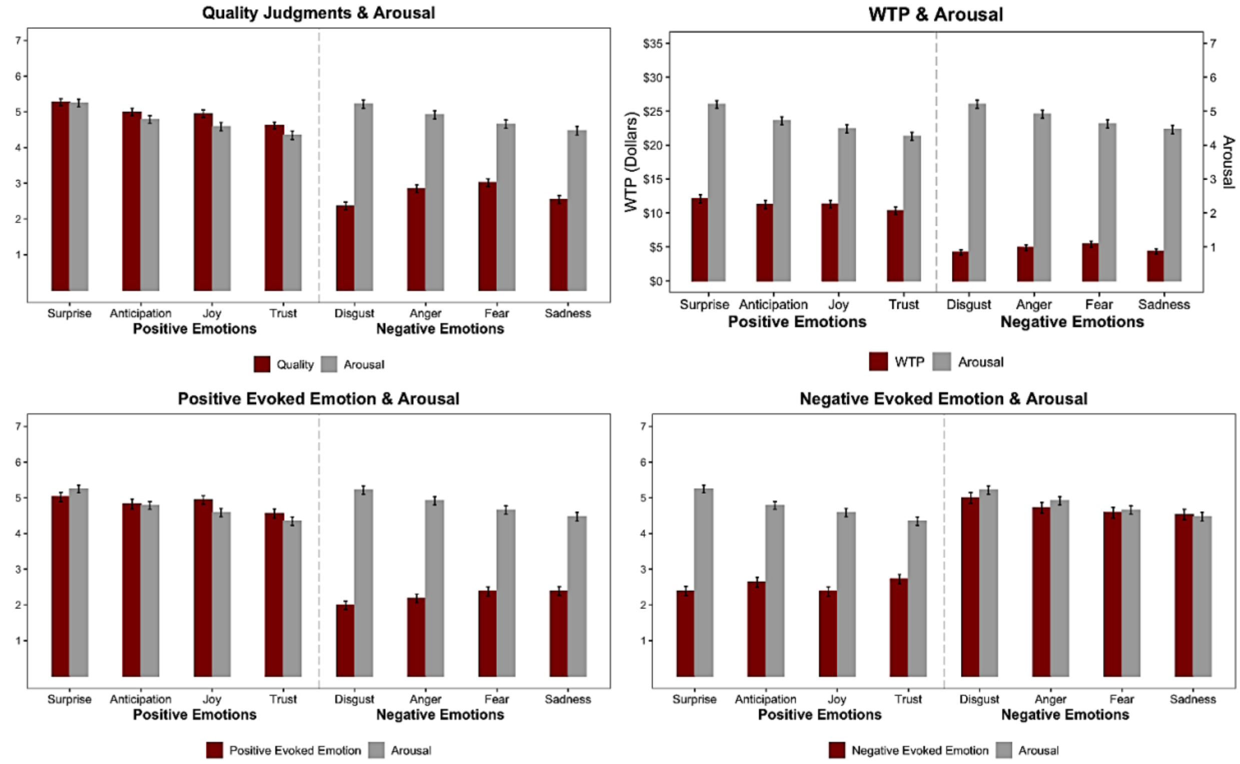
Quality Judgments. Next, we regressed perceived quality on a dummy variable for review valence, participants' arousal ratings, and their interaction, clustering standard errors by participant. We also included helpfulness, objectivity, reliability, and informativeness in the regression, which does not qualitatively change our main results (see Appendix 2.B for full regression table). We observed a significant interaction between valence and arousal ($\beta_{\text{Interaction}}$ = 0.29, t(798) = 7.17, p < .001; see Figure 2.8). For positive reviews, quality judgments were positively related to arousal ($\beta_{\text{Simple Slope}}$ = 0.26, t(798) = 9.26, p < .001). For negative reviews, quality judgments were negatively related to arousal, though the simple slope was not significant ($\beta_{\text{Simple Slope}}$ = -0.04, t(798) = -1.04, p = .298). A closer examination reveals this is due to the sadness condition, which had the lowest arousal of the negative reviews but only had the second-highest perceived quality. The other three negative emotions follow the predicted pattern.

WTP. We ran the same regression with WTP as the dependent variable and found the same results we observed for quality judgments. There was a significant valence-arousal interaction ($\beta_{\text{Interaction}}$ = 0.83, t(738) = 4.16, p < .001; see Figure 2.8), such that arousal was positively related to WTP for positive reviews ($\beta_{\text{Simple Slope}}$ = 0.63, t(738) = 3.87, p < .001) and negatively related to WTP for negative reviews ($\beta_{\text{Simple Slope}}$ = -0.20, t(738) = -1.56, p = .115). The average WTP in the sadness again did not follow the expected pattern, resulting in an insignificant simple slope for negative reviews.

Evoked Emotion. We ran the same regression and again replicated the interaction found in previous studies for both positive and negative evoked emotion (Positive Evoked: $\beta_{\text{Interaction}}$ = 0.52, t(798) = 10.16, p < .001, Negative Evoked: $\beta_{\text{Interaction}}$ = -0.51, t(798) = -8.07, p < .001). For positively valenced reviews, positive evoked emotion was greater for reviews with higher arousal ($\beta_{\text{Simple Slope}}$ = 0.41, t(798) = 10.72, p < .001). For negatively valenced reviews, negative evoked emotion was greater for reviews with higher arousal ($\beta_{\text{Simple Slope}}$ = 0.37, t(798) = 7.32, p < .001).

Moderated Mediation. We ran the same moderated mediation as in Studies 3, 4A, and 4B with one minor adjustment; the arousal moderator is now participant ratings instead of an experimental factor. The results we found replicate those in prior studies. For quality judgments, there was a significant effect of valence that was mediated by positive and negative evoked emotion and moderated by arousal. The indirect effects via evoked emotion become larger as arousal increases (Bootstrapped 95% CI's: Index of $\text{Moderated}_{\text{via Positive Evoked}}$ = [0.10, 0.15], Index of $\text{Moderated}_{\text{via Negative Evoked}}$ = [0.03, 0.07], see Appendix 2.B for full output). It is worth noting that the indirect effects and index of moderated mediation are smaller for the path via negative evoked emotion, again likely due to the sadness condition. We found the same result when using WTP as the dependent variable (Bootstrapped 95% CI's: Index of $\text{Moderated}_{\text{via Positive Evoked}}$ = [0.27, 0.46], Index of $\text{Moderated}_{\text{via Negative Evoked}}$ = [0.04, 0.19]; see Appendix 2.B for full output).

Figure 2.8: Study 5 Results



NOTE - All error bars are 95% confidence intervals (standard errors clustered by participant)

Discussion. Study 5 investigated the role of basic emotions in participants' judgments. We found that reviews associated with different basic emotions produce significant variation in arousal ratings. Additionally, the pattern of arousal ratings generally maps onto the patterns observed for quality judgments, WTP, and positive and negative evoked emotion. Thus, we replicate the same interactions found in Studies 3, 4A, and 4B by manipulating the basic emotion that the review text is associated with. We also replicated the results of the moderated mediation analyses.

While valence and arousal generally did a very good job at describing differences between these states, there was still value in knowing the specific basic emotion in terms of predicting quality judgments and WTP. Specifically, there seems to be something unique about sadness that weighs on purchase intentions, beyond its valence and arousal (note: this was true for all three sets of reviews used for stimulus sampling). One limitation of this study is the exact information in the reviews is not held constant, but the results provide additional insights into how different emotional states affect consumer purchase likelihood.

General Discussion

Six studies showed evidence highlighting the importance going beyond a single, unidimensional scale when measuring emotions. Study 1 provided initial evidence that evoked emotion has significant explanatory power in predicting quality judgments, even after controlling for expressed emotion. Evoked emotion was also found to be a stronger predictor of perceived quality. In Study 2, we used mixed reviews to demonstrate the value of measuring positive and negative emotion as separate constructs. Emotional mixed reviews that contained roughly equal amounts of positive and negative emotion were indistinguishable from bland mixed reviews when measuring valence on a single bipolar scale. These two types of reviews could only be differentiated with unipolar scales, which is important because they produced different patterns of quality judgments. Studies 3, 4A, and 4B showed the importance of accounting for arousal, as the arousal of reviews tended to amplify the effect of valence on purchase decisions. They also provided evidence supporting positive and negative evoked emotion as significant mediators of that effect. Lastly, Study 5 explored the role of basic emotions in product reviews and their effect on arousal, evoked emotion, and product evaluations. Arousal again amplified the effect of review valence on evoked emotion, quality judgments, and WTP. Taken together, these findings make several theoretical and practical contributions.

Theoretical Contributions

This work makes a significant contribution to the literature on consumer reviews by demonstrating the importance of considering evoked emotion. We found evoked emotion to be a strong predictor of quality judgments, even after controlling for the emotion expressed by the author. Furthermore, we identify positive and negative evoked emotion as psychological mechanisms through which valence and arousal affect consumer decisions. In our mediation models from Studies 3, 4A, and 4B, the indirect effects suggest that evoked emotion functions as a core mechanism by which expressed emotion impacts product evaluations. This was especially apparent at higher levels of arousal. This model improves our understanding of how reactions to product reviews influence consumer judgments and behaviors.

We also contribute to the literature on the independence of positive and negative emotion. Our results most closely align with the theory that the level of independence between two is context dependent (Dejonckheere et. al., 2021). In the context of consumer reviews, bipolar scales for valence may work just as well as unipolar scales when the emotions in the review are unambiguously positive or negative. However, when reviews are mixed, as in Study 2, the independence of the two becomes important and treating valence as bipolar would result in an ambiguous midpoint. Additionally, one cannot simply account for the Study 2 results by using a bipolar scale for valence along with a measure of arousal, as such an account would predict a difference between high- and low-arousal reviews that we did not observe.

Given these findings, we believe it is important to measure positive and negative emotions separately, at least in the context of product reviews.

This work also adds to the body of research on categorical models of emotion and basic emotions. While we cannot answer the question of whether basic emotions exist, words associated with emotions that are often considered basic do produce predictable differences in consumer judgments. If there are basic emotions, the results of Study 5 suggest that the valence-arousal framework describes the differences between those emotions quite well (though not perfectly). Basic emotions that were higher in arousal tended to be preferred for positive reviews, while the opposite was true for negative reviews. Studies 4A and 4B produced similar conclusions by explicitly giving participants a discrete emotional state that varied in terms of valence and arousal. These results lead us to agree with Russell (1980) that the categorical models of emotion compliment, rather than contradict, dimensional models of emotion. While there may be value to classifying emotions into discrete categories, measuring valence and arousal offers a fairly accurate and parsimonious representation of emotion in our studies.

Marketing Implications

Our results also have consequential implications for marketers and firms. First, they should account for arousal in consumer reviews. Sentiment analysis tools are popular in marketing but often give firms an overly simplistic metric of how positive or negative a review is. That method of measuring emotion may be useful, but additional insights can be gleaned by considering arousal as well. Without accounting for arousal, there is a potential confound when attempting to interpret the output of a unidimensional sentiment analysis tool. This can lead to firms missing out on important conclusions that could improve sales and/or customer relations. For example, if a firm notices very high sentiment scores in their reviews but muted sales, it could be that most of the positive reviews are low in arousal while the negative reviews are high in arousal. Without understanding the role of arousal, firms may draw incorrect or incomplete conclusions about how the valence of product reviews relates to consumer behavior.

A second implication is that researchers and marketers should consider the role evoked emotion plays in driving responses to consumer reviews. While tools like sentiment analysis are adept at providing ratings for the valence expressed by people writing reviews, they are not typically designed to measure how the reviews make the reader feel. Our results suggest that evoked emotion is a significant predictor of consumer judgments, even more so than expressed emotion. While we find significant, positive correlations between evoked and expressed emotion, they do not appear to be synonymous. Our results suggest the two diverge more as reviews become more emotional. Thus, using reviews to predict behavior of prospective customers may be less accurate for products with more emotional reviews, as measuring the emotions in those reviews may not yield a clear picture of the emotions felt by the reader.

Lastly, firms may benefit differentially from measuring valence with unipolar (vs. bipolar) scales. For products that produce extreme opinions (i.e., lots of one- and five-star reviews), measuring valence on a bipolar scale may work well. For products that have more mixed opinions (i.e., more two-, three-, and four-star reviews), firms may benefit from measuring positive and negative emotion separately. We find mixed reviews that are more emotional produce greater heterogeneity in perceived product quality. This finding has key marketing implications as firms choose pricing and promotion tactics. For products with highly emotional reviews, there may be a greater benefit to price discrimination and targeted promotions due to greater heterogeneity in quality judgments. Future research could further explore the root cause of that heterogeneity and how marketers can target different parts of the perceived quality distribution.

Limitations and Future Directions

While our work has important theoretical and marking implications, there are several limitations that present opportunities for future research. First, holding everything constant is extremely difficult when dealing with language, creating potential confounds in some of the study designs. Study 1 used real reviews to provide high external validity, but the reviews can differ in many ways. While we included several control measures we believed could be related to perceived quality, we cannot be sure we measured and controlled for every key covariate. Also, while we tried to use a large, representative set of products, there may be certain products or categories that behave differently. Future research could use more data-driven approaches to identify a richer set of features in reviews that predict product perceptions. There are some NLP models that measure both valence and arousal (Mohammad & Turney, 2013; Mohammad, 2016) or evoked emotion (Chang et. al., 2016), which could be useful in mining for additional predictors.

Study 2 used 3-star reviews, which are generally the least common rating (Hu, Zhang, & Pavlou, 2009). Thus, Study 2 is somewhat lacking in external validity. However, 3-star reviews do exist for virtually every product, and they provide an ideal setting to test the effects of mixed reviews in general. This is valuable because reviews do often have both pros and cons (Lu, Qiu, & Wang, 2021) and some websites explicitly display positive and negative reviews together. Future research could further explore the effect of mixed emotions in reviews by manipulating positive and negative emotions across reviews to see if that has the same effect as manipulating it within reviews. It also may be interesting to explore cases where there is a mixture of positive and negative emotion, but one is more intense than the others. One could try to identify how changing the relative mix of positive and negative emotions affects the benefit of using unipolar scales.

Studies 3 – 5 all addressed the effect of valence and arousal on product evaluations. One limitation of Study 3 is the arousal manipulation may be confounded with intensity. Studies 4A and 4B address that limitation but the manipulation may be less externally valid. Study 5 varies emotion in a way that also varies the specific pieces of information in the review, and we cannot be sure we controlled for all relevant covariates. Despite individual shortcomings, Studies 3 – 5 do provide convergent evidence of a valence-arousal interaction in evoked emotion and product evaluations.

Another potential limitation of these studies could be that we limit our analysis to valence and arousal. We had theoretical reasons for doing so (Barrett & Russell, 1999), but future research could explore adding additional dimensions. For example, future work could further investigate the role of intensity in product evaluations. Studies 4A and 4B hold intensity constant to prevent confounding, but do not provide insight into the effect of intensity. We ran one exploratory study that did experimentally manipulate valence, arousal, and intensity, but the results were largely uninformative regarding the intensity factor (see Appendix 2.A for details). This presents a fruitful opportunity for future work to study how intensity might moderate the effects of valence and arousal on consumer emotions and behavior.

Lastly, a limitation of our studies is they used stylized stimuli. We used a somewhat limited set of products and all studies used hypothetical scenarios. Future work could examine the effects of valence and arousal across different dimensions of product characteristics (e.g., material-experiential, hedonic-pragmatic, price, etc.). Future research could also test our conclusions using real purchase behavior, either via consequential experiments or field data. This could further corroborate our findings and potentially give quantitative estimates for how much firms can benefit from considering arousal, evoked emotion, and/or the independence of positive and negative emotion. It would also answer whether these effects exist "in the wild," where there are infinitely more factors that influence purchase decisions.

One other avenue for future work is examining other factors that affect the correlation between evoked and expressed emotion. Study 2 finds that correlation is lower when reviews are more emotional. Supplemental Study 1, which varied the intensity along with the valence and arousal, found suggestive evidence that intensity may affect evoked emotion differently depending on the valence (see Appendix 2.A for full details). In that study, evoked and expressed emotion were directionally most similar at higher intensities for positive reviews, but most similar at low intensities for negative reviews. Additionally, in our main studies, the effect of arousal was generally weaker for negative reviews than positive reviews throughout all our studies. Taken together, this suggests a potential asymmetry where the relationship between evoked and expressed emotion may vary by valence. Future research could further probe this asymmetry to gain additional insight into how positive and negative reviews affect consumer emotions differently.

Conclusion

Online shopping and advances in text analysis have made researchers and marketers quite interested in the emotional content of product reviews. However, the vast majority of work in this space focuses solely on the valence of the review. We drew on long-standing theories of emotion to show that arousal, in conjunction with valence, can significantly influence consumers' emotions, judgments, and choices. We also highlighted the need to consider the emotion experienced by the reader, as those feelings significantly affect consumer behavior and are not synonymous with the emotion expressed by the writer. Lastly, we demonstrated the need to consider positive and negative emotion separately, particularly for mixed reviews. Integrating these findings to existing research on emotion in product reviews can allow researchers and marketers to gain a better understanding of how reviews impact consumer behavior.

Supplemental Material for Chapter 1

Appendix 1.A: Supplemental Study 1 – Absolute Deviation and Willingness to Pay

The studies in the main text examined the effect of a review's absolute deviation from the mean rating on review helpfulness and search behavior. In Supplemental Study 1 (hereafter Study S1), we test whether this relationship between deviation from the mean and helpfulness translates into an effect on purchase behavior. If a review is more (less) helpful, it should have a greater (lesser) impact on purchase intentions. Given absolute deviation affects helpfulness, Study S1 tests this by examining participants' willingness to pay (WTP) for a book after reading a review while varying the review and mean rating.

Experimental Design and Procedure

Study S1 uses a 3(review rating: 2, 3, or 4)[13] x 3(mean rating: 2, 4, or none) between-subjects design. Participants read one book review and then gave their WTP. We included a control condition where we showed no mean rating to examine the WTP based on the review alone. For participants who see a review that matches (does not match) the mean rating, the effect on WTP should be amplified (attenuated).

[13] The 3-star review rating condition was exploratory.

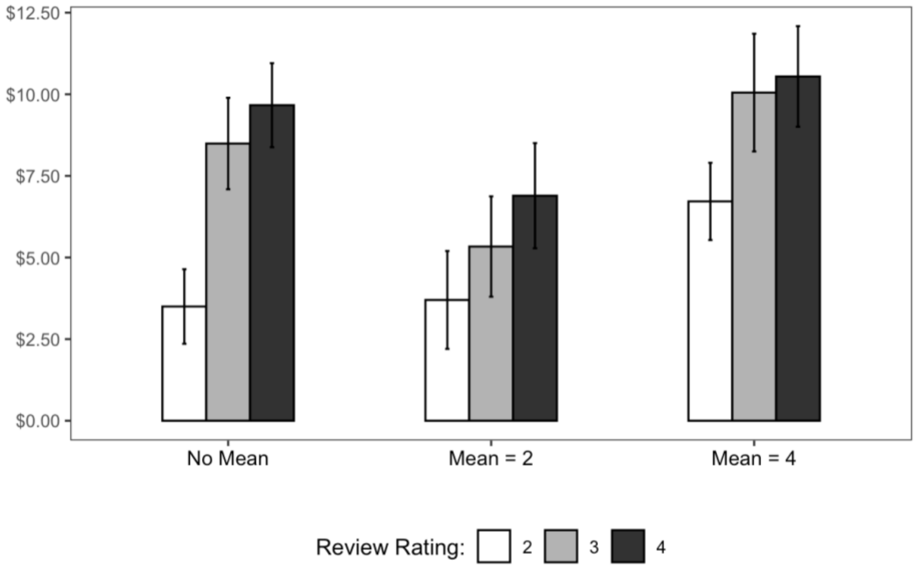
Participants

Four hundred and fifty-one participants completed the survey on Prolific (M_{age} = 35, 48% female).

Results

Figure 1.A1 shows the WTP for each condition. Across the levels of the mean rating factor, we calculated the difference in WTP among those who saw the 4-star review and the WTP among those who saw the 2-star review. If absolute deviation causes reviews to impact WTP more when they are equal to the mean rating and less when they are far from the mean rating, that difference should be greater when the mean is two or four compared to when there is no mean. However, we do not see evidence of that, as the difference is actually largest when there is no mean ($WTP_{4\text{-star}} - WTP_{2\text{-star}}$: $M_{No\ Mean}$ = 6.17, $M_{Mean\ =\ 2}$ = 3.19, $M_{Mean\ =\ 4}$ = 3.83).

Figure 1.A1: Study S1 WTP Results



Discussion

In Study S1, we examined WTP to study the effect of a review's absolute deviation from the mean on WTP. While WTP was generally higher when the review or mean rating was higher (as expected), we did not see evidence that a review's effect on WTP is amplified when the review rating equals the mean. If the mean being two caused the two-star review to have a more negative effect on product perceptions, and the mean being four caused the four-star review to have a more positive effect on product perceptions, we would expect the WTP gap between the two- and four-star reviews to be lowest in the condition with no mean rating. We did not find this. Future research should further examine the roles of a review's deviation from the mean and its helpfulness on eventual purchase behavior.

Appendix 1.B: Supplemental Study 2 – Direction of Confirmation Bias

The results of the studies in the main text are consistent with confirmation bias in judgments of review helpfulness. However, confirmation bias is not a single phenomenon, it is a collection of similar behavioral biases (Klayman, 1995; Nickerson 1998). The studies in main text, provide evidence for both backward- (Study 3) and forward-looking (Studies 4A, 4B, and 5) confirmation biases. In Supplemental Study 2 (hereafter Study S2), we further assessed the direction of the confirmation bias we find in our studies.

Additionally, Study S2 includes a control condition where the review's text is presented with no star rating attached. In previous studies, we cannot discern whether the negative relationship between deviation and helpfulness is due to reviews close to the mean being especially helpful, reviews far from the mean being especially unhelpful, or both. To resolve this, Study S2 compares helpfulness at each level of absolute deviation to this control.

Experimental Design and Procedure

This study used a 2(average rating: 2 or 4) x 2(order: review first or average first) x 3(star rating of review: 1, 5, or the average rating) + 1(control) between-subjects design. Unlike earlier studies where the summary information and the review appeared on the same screen, the two appeared sequentially in this study. We counterbalanced the order such that one condition saw the summary information on the first screen and the review on the second screen, while the other saw the reverse. In the control condition, participants saw a review but never saw an average product rating. Control participants randomly saw a one-, two-, four-, or five-star review. Every participant judged the helpfulness of a single review.

If our results were solely driven by forward-looking confirmation bias, showing the review before the star rating should attenuate the effect because the mean is not there to bias one's processing of the review. If backward-looking confirmation bias is involved, we would still expect a significant relationship between deviation and helpfulness, as there is still the opportunity to retroactively reevaluate the review after seeing the mean. We included conditions where the summary information appeared first to assess whether our prior results hold when the summary information and the review are presented sequentially.

We omitted certain conditions in this study that appeared in prior studies to simplify the design and increase the statistical power of our tests. We focused on more extreme deviation levels, as that is where we observed the largest effects (e.g., we omitted a mean rating of three stars). We also omitted the product and distribution factors, as those have not meaningfully impacted our results.

Participants

Three hundred and fifty-two participants completed the survey on Prolific (M_{age} = 31, 52% female). Four participants were excluded due to a memory check failure, leaving 348 valid completions.

Results

To investigate the nature of the confirmation bias, we regressed review helpfulness on the terms listed in Table 1.B1 (excluding participants in the control condition). We standardized all continuous variables and centered star rating at 3. Presentation order is deviation coded categorical variable that takes on a value of 0.5 for the condition that saw the review first and -0.5 for the condition that saw the mean rating first. Table 1.B1 shows the results.

Table 1.B1: Study S2 Helpfulness Regression

Predictors	Estimate	t	p
Intercept	0.06	0.84	.402
Deviation From Mean	**-0.30**	**-4.11**	**< .001**
Order [Review First]	-0.12	-1.14	.254
Star Rating	-0.03	-0.53	.595
Deviation From Mean*Order [Review First]	-0.22	-2.04	.042
Deviation From Mean*Star Rating	0.10	1.81	.071
Observations	297		
R^2	0.19		
R^2 adjusted	0.18		

NOTE - Continuous variables standardized, categorical variables deviation coded

The relationship between deviation and helpfulness was present regardless of presentation order, which suggests the presence of backward-looking confirmation bias.

Another important part of this study was the control condition. We compared the helpfulness reported by participants in the control conditions to those of participants at each level of absolute deviation (see Figure 1.B1). To do this, we ran an ANOVA with review helpfulness as the dependent variable and deviation condition, star rating of the review, and their interaction as predictors. Deviation condition is a categorical variable for deviation from the mean with four levels: zero, one, three, and control. The results appear in Table 1.B2.

Table 1.B2: Study S2 Helpfulness ANOVA

	SS	df	F	p
Intercept	158.55	1	54.93	< .001
Deviation Condition	**73.04**	**3**	**8.43**	**< .001**
Star Rating	9.83	1	3.40	0.066
Interaction	17.05	3	1.97	0.118
Residuals	981.41	340		
				Type III Sum of Squares

Examining pairwise comparisons, we observed that perceived helpfulness of reviews with no available mean rating is comparable to when the review is one star away from the mean. All pairwise comparisons between the four mean rating conditions (collapsed across other experimental factors) were significant except the comparison between the control condition and the condition with a deviation of one. Having a deviation of zero (three) pushed perceived helpfulness above (below) that baseline.

Figure 1.B1: Study S2 Helpfulness Ratings

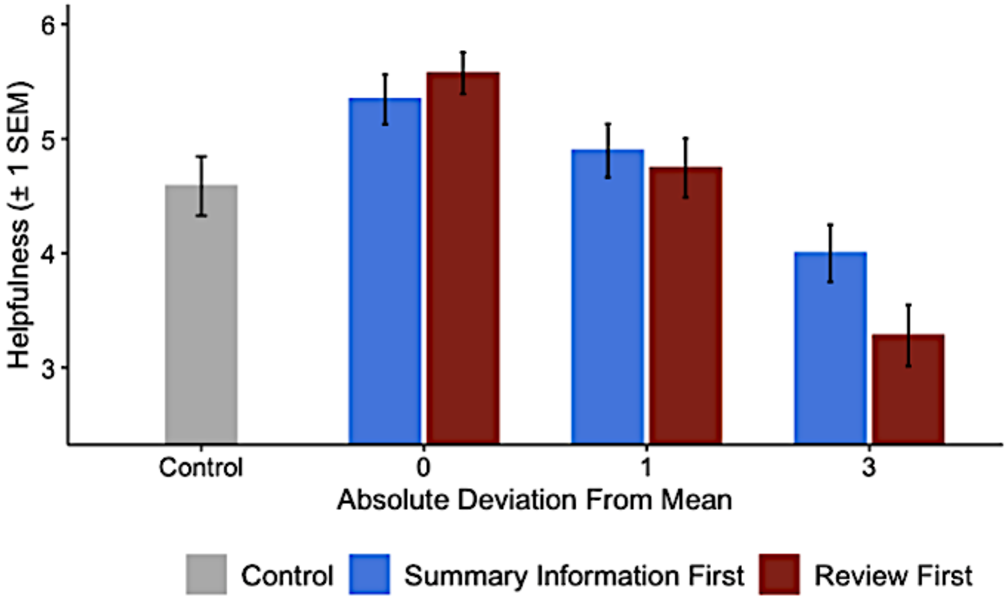


Table 1.B3: Study S2 Helpfulness Pairwise Comparisons

Deviation From Mean	Estimate	S.E.	t	p
Zero - Control	0.86	0.29	2.93	.004
One - Control	0.22	0.29	0.76	.445
Three - Control	-0.96	0.29	-3.3	.001
Zero - One	0.64	0.24	2.63	.009
Zero - Three	1.82	0.24	7.53	< .001
One - Three	1.19	0.24	4.95	< .001

Discussion

This study provided a richer understanding of the patterns we observed in prior studies. First, Study S2 suggests the relationship between deviation and helpfulness is influenced by both forward- and backward-looking confirmation biases. Unexpectedly, the pattern was stronger when the review was presented first. A possible explanation could be that, because the mean rating can act as a reference point, placing it closer to the helpfulness judgement may produce a stronger effect due to recency. However, the important finding is helpfulness was negatively related to absolute deviation regardless of whether participants saw the mean before or after reading the review. Additionally, the control condition provided additional information on what drives the negative relationship between deviation and helpfulness. Helpfulness in the control condition fell in between the two extremes in absolute deviation, suggesting the negative relationship is not driven solely by small- or large-deviation reviews. Both of these results shed light on the cognitive processes underlying that relationship.

Appendix 1.C: Cue Weighting from Study 3

Consumers update their beliefs about products based on the signals they attend to when reading reviews. An additional goal of Study 3 was to explore how they use those signals. When a consumer reads a review, there are two main cues they can use to update their beliefs, the star rating and the text. Study 3 directly probed how important these cues are to participants when forming their helpfulness judgements.

Cue-Weighting Procedure

In addition to the helpfulness and belief updating questions from Study 3, participants also answered the following question: “What was the most helpful part of the review?” (1 = definitely star rating, 7 = definitely text; counterbalanced).

Results

We ran the same regression from Study 3 on the importance of the text (relative to the star rating) in participants’ helpfulness judgments (see Table 1.C1).

Table 1.C1: Study 3 Cue Weighting Regression

	Text > Star		
Predictors	Estimate	t	p
Intercept	0.00	0.00	.997
Deviation from Mean	**-0.09**	**-5.82**	**< .001**
Star Rating	-0.06	-2.75	.006
Distribution [Mean ≠ Mode]	0.01	0.14	.887
Product [Book]	-0.19	-3.78	< .001
Deviation from Mean*Star Rating	-0.08	-3.45	.001
Deviation from Mean*Distribution [Mean ≠ Mode]	0.02	0.50	.619
Deviation from Mean*Product [Book]	0.01	0.32	.748
Star Rating*Product [Book]	0.36	10.87	< .001
Observations	3150		
R^2	0.06		
R^2 adjusted	0.06		

NOTE- Continuous variables standardized, categorical variables deviation coded

Discussion

As deviation from the mean increased, participants put less weight on the review's text and more weight on the review's rating when forming helpfulness judgments. This suggests deviation from the mean might also impact how much attention consumers pay to the text when reading the reviews.

Appendix 1.D: Comparing the Book and Blender

The goal of this study was to attempt to understand the key differences between the blender and book that could explain why we observed this difference in positivity and negativity bias. We drew on prior literature to create a list of dimensions along which to compare the blender and book. Some research has found consumers rely on consumer reviews more for material products than for experiential products (Dai, Chan, & Mogilner, 2020). There is also work suggesting that negative reviews are more helpful than positive reviews for hedonic goods, but not for pragmatic goods (Sen & Lerman, 2007). Many of these findings from prior literature suggest reviews are more helpful when they are viewed as depicting a product's inherent quality and less helpful when viewed as more idiosyncratic to a reviewer or experience. Thus, we used several measures to investigate how the blender and book differ on dimensions of this sort. Lastly, prior work has found consumers' goals when evaluating products influence the types of reviews they find helpful. Consumers with promotion goals tend to find positive reviews more helpful and consumers with prevention goals tend to find negative reviews more helpful (Higgins, 1987; Zhang, Craciun, & Shin, 2010). Therefore, we also tested whether the blender and book differ in their associations with promotion and prevention goals.

Experimental Design and Procedure

This study had 2(product: blender, book) between-subjects conditions. Participants then responded, in a random order, to several scale measures regarding the product to which they were assigned. We used previously validated scales to measure how hedonic (vs. pragmatic; Voss, Spangenberg, & Grohmann, 2003), objective (vs. subjective; Loureiro, Garcia-Marques, & Wegener, 2020), complex (vs. simple; Loureiro, Garcia-Marques, & Wegener, 2020), experiential (vs. material; Caprariello & Reis, 2013), and promotion-focused (vs. prevention-focused; Zhang, Craciun, & Shin, 2010) the products were.

Participants

One hundred and thirty participants completed the survey on Prolific (M_{age} = 32, 58% female).

Results

The results for each scale measure appear in Table 1.D1. The only measures where the blender and book did not differ significantly were the pragmatic subscale of the hedonic/pragmatic scale and the promotion subscale of the regulatory focus scale.

Table 1.D1: Product Scale Measure Results

	Blender		Book		Blender - Book Difference	
	Mean	**(S.D.)**	**Mean**	**(S.D.)**	**t-statistic**	**p-value**
Hedonic Rating	4.15	(1.19)	5.56	(1.31)	-6.43	< .001
Pragmatic Rating	5.95	(0.88)	6.11	(1.07)	-0.92	0.36
Experiential Rating	2.44	(1.69)	4.13	(1.80)	-5.49	< .001
Material Rating	6.18	(1.26)	5.34	(1.63)	3.28	0.001
Objectivity Rating	4.68	(1.29)	2.70	(1.14)	9.27	< .001
Complexity Rating	3.14	(1.26)	4.31	(1.74)	-4.11	< .001
Promotion Focus Rating	5.39	(0.94)	5.28	(0.81)	0.75	0.45
Prevention Focus Rating	3.86	(1.03)	3.23	(0.78)	3.93	< .001

Discussion

Results from this study revealed that, relative to the blender, the book was viewed as more hedonic, experiential, and subjective. These attributes generally relate to reviews that are more idiosyncratic to the reviewer. These results, combined with the results from studies in the main text, suggest positive (negative) reviews are more helpful for products with more (less) idiosyncratic reviews.

The difference in prevention focus invoked by the two products could also have played a role, although there was no significant difference in promotion focus. More work is needed to discern if regulatory focus helps explains the difference in positivity/negativity bias across these types of products.

Appendix 1.E: Additional Analyses

Study 1B

OLS Regression Ignoring Star Rating

	Helpfulness		
Predictors	Estimate	t	p
Intercept	0	0	.998
Deviation from Mean	-0.32	-5.78	< .001
Observations	294		
R^2	0.10		
R^2 adjusted	0.10		

NOTE - Variables standardized

OLS Regression Using Signed Deviation[14]

	Helpfulness		
Predictors	Estimate	t	p
Intercept	0.00	-0.01	.995
Absolute Deviation	-0.32	-5.79	< .001

[14] In this and all studies the effect of deviation is negative regardless of the sign of the deviation.

Deviation Sign	-0.01	-0.10	.922
Absolute Deviation*Deviation Sign	-0.10	-1.28	.201
Observations	294		
R^2	0.11		
R^2 adjusted	0.10		

NOTE - Continuous variables standardized

Study 2

OLS Regression with all 2- and 3-way Interactions

	Helpfulness		
Predictors	Estimate	t	p
Intercept	4.68	177.62	< .001
Deviation from Mean	-0.39	-14.84	< .001
Star Rating	-0.21	-7.06	< .001
Distribution [Mean ≠ Mode]	0.03	0.48	.630
Product [Book]	-0.63	-12.01	< .001
Total Reviews [High]	-0.02	-0.46	.648
Deviation from Mean*Star Rating	-0.07	-2.68	.007
Deviation from Mean*Distribution [Mean ≠ Mode]	0.07	1.29	.198
Deviation from Mean*Product [Book]	-0.02	-0.44	.662
Deviation from Mean*Total Reviews [High]	0.01	0.17	.861
Star Rating*Distribution [Mean ≠ Mode]	0.04	0.63	.528
Star Rating*Product [Book]	0.49	8.24	< .001
Star Rating*Total Reviews [High]	0.01	0.16	.871
Distribution [Mean ≠ Mode]*Product [Book]	0.05	0.46	.648

Distribution [Mean ≠ Mode]*Total Reviews [High]	-0.07	-0.62	.537
Product [Book]*Total Reviews [High]	0.08	0.73	.466
Deviation from Mean*Star Rating*Distribution [Mean ≠ Mode]	-0.04	-0.67	.501
Deviation from Mean*Star Rating*Product [Book]	0.27	4.98	< .001
Deviation from Mean*Star Rating*Total Reviews [High]	-0.01	-0.24	.811
Deviation from Mean*Distribution [Mean ≠ Mode]*Product [Book]	0.07	0.65	.513
Deviation from Mean*Distribution [Mean ≠ Mode]*Total Reviews [High]	-0.21	-1.99	.047
Deviation from Mean*Product [Book]*Total Reviews [High]	-0.07	-0.64	.519
Star Rating*Distribution [Mean ≠ Mode]*Product [Book]	-0.08	-0.78	.435
Star Rating*Distribution [Mean ≠ Mode]*Total Reviews [High]	0.06	0.53	.598
Star Rating*Product [Book]*Total Reviews [High]	-0.04	-0.33	.739
Distribution [Mean ≠ Mode]*Product [Book]*Total Reviews [High]	-0.21	-0.99	.324
Observations	3603		
R^2	0.15		
R^2 adjusted	0.15		

NOTE - Continuous variables standardized, categorical variables deviation coded

Study 4

OLS Regression for Helpfulness

	Helpfulness		
Predictors	Estimate	t	p
Intercept	0.09	1.21	.225
Mean rating = 4	-0.39	-3.51	< .001
Review rating = 4	-0.18	-1.68	.093
Interaction	0.76	4.90	< .001
Observations	645		
R^2	0.045		
R^2 adjusted	0.041		

NOTE - Continuous variables standardized

Mediation Model Regressions (Moderator on b- and c-paths)

Bootstrap Results for Mediation Model Regressions		
Outcome Variable: Relative Positivity		
	Estimate	**95% CI**
Intercept	-0.58	[-0.77, -0.39]
Average Rating [=4]	0.60	[0.34, 0.86]
Outcome Variable: Helpfulness		
	Estimate	**95% CI**
Intercept	4.89	[4.63, 5.15]
Average Rating [=4]	-0.49	[-0.85, -0.13]
Relative Positivity	-0.18	[-0.29, -0.08]
Review Rating [=4]	-0.20	[-0.56, 0.18]
Average Rating [=4]*Review Rating [=4]	1.06	[0.57, 1.54]
Relative Positivity*Review Rating [=4]	0.21	[0.08, 0.35]

	Estimate	**95% Bootstrapped CI**
Indirect Effect (2-star reviews)	-0.11	[-0.19, -0.04]
Indirect Effect (4-star reviews)	0.02	[-0.03, 0.08]
Index of Moderated Mediation	0.13	[0.04, 0.24]

Mediation Model Regressions (Moderator on a-, b-, and c-paths)

Bootstrap Results for Mediation Model Regressions		
Outcome Variable: Relative Positivity		
	Estimate	**95% CI**
Intercept	-0.38	[-0.63, -0.12]
Average Rating [=4]	0.62	[0.25, 0.98]
Review Rating [=4]	-0.41	[-0.77, -0.38]
Average Rating [=4]*Review Rating [=4]	-0.04	[-0.57, 0.48]
Outcome Variable: Helpfulness		
	Estimate	**95% CI**
Intercept	4.89	[4.63, 5.15]
Average Rating [=4]	-0.49	[-0.85, -0.13]
Relative Positivity	-0.18	[-0.29, -0.08]
Review Rating [=4]	-0.20	[-0.56, 0.18]
Average Rating [=4]*Review Rating [=4]	1.06	[0.57, 1.54]
Relative Positivity*Review Rating [=4]	0.21	[0.08, 0.35]

	Estimate	**95% Bootstrapped CI**
Indirect Effect (2-star reviews)	-0.11	[-0.21, -0.04]
Indirect Effect (4-star reviews)	0.02	[-0.04, 0.08]
Index of Moderated Mediation	0.13	[0.03, 0.25]

Study 5

Linear Mixed-Effects Regression for Helpfulness

	Helpfulness		
Predictors	Estimate	t	p
Intercept	0.00	0.02	.988
Deviation From Mean	-0.23	-15.33	< .001
Star Rating	-0.20	-11.38	< .001
Distribution [Mean ≠ Mode]	0.02	0.35	.723
Product [Blender]	0.14	1.86	.063
Product [Painting]	0.06	0.82	.414
Product[Trash Can]	0.30	3.87	< .001
Deviation From Mean*Star Rating	-0.02	-1.24	.215
Deviation From Mean*Distribution [Mean ≠ Mode]	-0.12	-4.07	< .001

Deviation From Mean*Product [Blender]	-0.03	-0.51	.609
Deviation From Mean*Product [Painting]	-0.05	-0.88	.381
Deviation From Mean*Product[Trash Can]	0.22	4.26	< .001
Star Rating*Product [Blender]	-0.54	-10.34	< .001
Star Rating*Product [Painting]	0.07	1.31	.191
Star Rating*Product[Trash Can]	-0.15	-2.83	.005
Random Effects			
σ^2	0.68		
$\tau_{00\ id}$	0.15		
ICC	0.19		
N_{id}	600		
Observations	3000		
Marginal R^2	0.17		
Conditional R^2	0.33		

NOTE - Continuous variables standardized, Categorical variables deviation coded

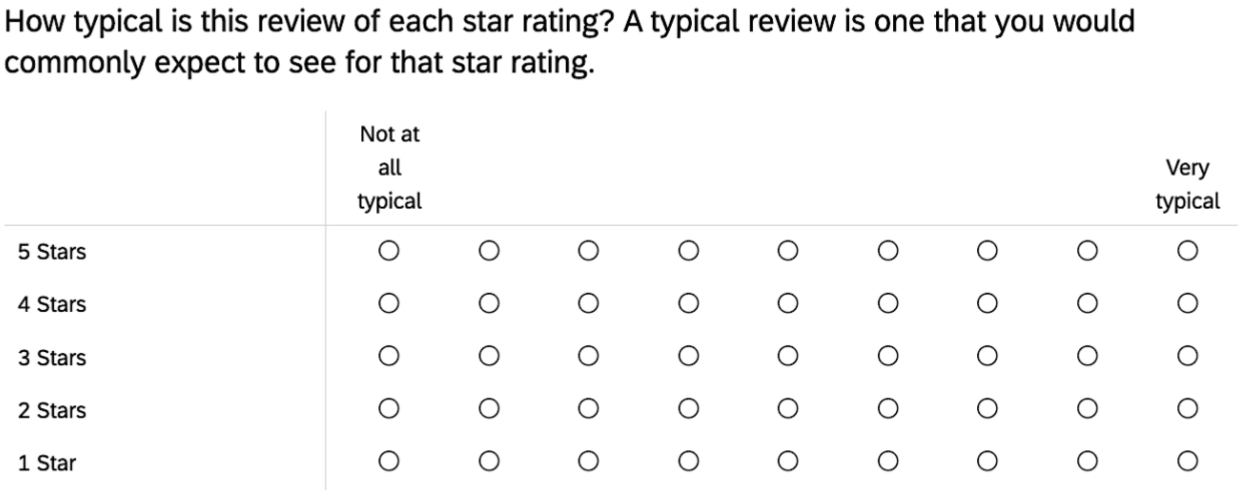
Appendix 1.F: Norming Study for Reviews

The reviews used in our studies are adapted from Amazon reviews (edited to remove personal information, reference to specific brands/products, typos, etc.). We wanted to ensure the text of the review appropriately matches the star rating we paired it with. For example, we clearly do not want a negative review to be associated with a five-star rating. Thus, to select reviews, we wanted reviews that were highly typical of one star rating and atypical of other star ratings. In other words, we want our five-star reviews to be reflective of a typical five-star review, our four-star reviews to be reflective of a typical four-star review, etc.

Experimental Design and Procedure

This study used a 5(Review Set: A, B, C, D, or E; between) x 5(True Review Rating: 1, 2, 3, 4, or 5 stars; within) mixed design. All participants saw five reviews, one from each star rating (importantly, they did not know this). Additionally, the star rating of the review was not displayed to participants. After reading the text of a given review, participants rated how typical the review was for a 5-star review, 4-star review, etc. (see Figure 1.F1).

Figure 1.F1: Typicality Question for Norming Study



How typical is this review of each star rating? A typical review is one that you would commonly expect to see for that star rating.

	Not at all typical								Very typical
5 Stars	○	○	○	○	○	○	○	○	○
4 Stars	○	○	○	○	○	○	○	○	○
3 Stars	○	○	○	○	○	○	○	○	○
2 Stars	○	○	○	○	○	○	○	○	○
1 Star	○	○	○	○	○	○	○	○	○

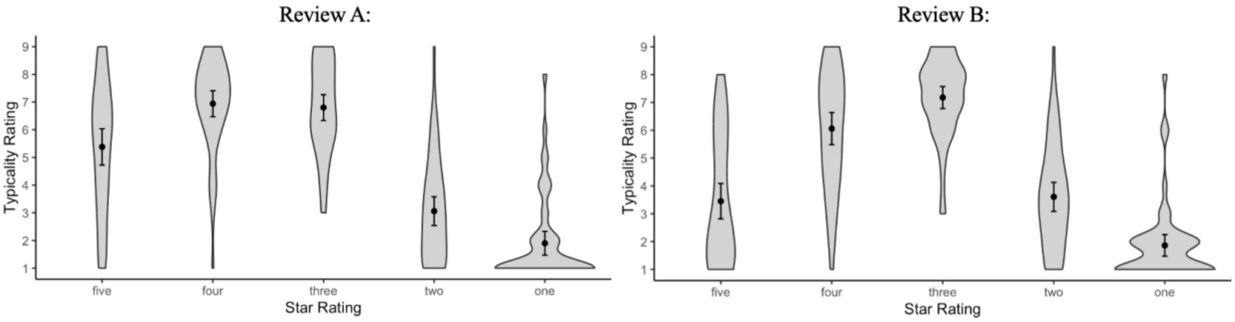
Participants

Five hundred and eighty-one participants completed the survey on Prolific (M_{age} = 32, 49% female). Eighty-one participants were excluded due to a memory check failure, leaving 500 valid completions.

Results

The supplemental files on OSF[15] contain a document that shows the typicality results by star rating for every review. We chose reviews that had relatively high typicality for one of the star ratings and relatively low typicality for the four other ratings. Figure 1.F2 shows an example of how we selected a three-star review. Review A shows a review that participants thought could be seen as a three- or four-star review, whereas Review B shows a review that is more unequivocally seen as three stars. Thus, we chose Review B.

Figure 1.F2: Two Potential Three-Star Reviews



[15] https://osf.io/kum3b/?view_only=1ce7516de70b4cd294bd4b83f2286e30

Discussion

This norming study allowed us to choose appropriate reviews for studies where we varied the text of the review with the star rating. In other studies, we held the text constant. Both procedures provided converging evidence that review helpfulness decreases as its rating strays further from the mean rating.

Supplemental Material for Chapter 2

Appendix 2.A: Supplemental Study 1 – Intensity, Evoked Emotion, and Purchase Behavior

Studies 4A and 4B attempt to hold intensity of the emotion constant by using discrete emotional states, allowing us to isolate the joint effect of valence and arousal. Supplemental Study 1 (hereafter Study S1) builds on those by varying the intensity of the emotion, along with the valence and arousal. In Study 2, we found evoked and expressed emotion diverged more for reviews that were more emotional. Thus, we hypothesized higher levels of intensity may cause evoked and expressed emotion to diverge more. This would be consistent with prior research that finds overly positive emotional language can backfire in certain contexts (Rocklage & Fazio, 2020). Given we have found evoked emotion to be a significant mediator of the effect of valence on product evaluations, we hypothesize, perhaps counterintuitively, that high levels of intensity would improve product evaluations for negative reviews and worsen them for positive reviews. We also include a medium-intensity condition for exploratory purposes. Prior research has found a quadratic relationship between a review's arousal and perceived helpfulness (Yin, Bond, & Zang, 2017). It is possible that increasing intensity initially amplifies our prior effects before backfiring when it becomes too intense.

Method

Participants. We recruited 1,200 participants online through Prolific (M_{age} = 37.62, SD_{age} = 12.85, 49% female). We excluded one participant who failed a memory check, leaving 1,199 valid completions. For analyses involving willingness to pay (WTP), we excluded 36 outliers via Tukey's Rule, leaving 1,163 participants for those analyses.

Design and Procedure. This study used a design almost identical to that of Studies 4A and 4B with one change; we added an additional between-subjects factor to vary intensity. Intensity could be low ("I was [emotion] …"), medium ("I was extremely [emotion] … !"), or high ("I was EXTREMELY [EMOTION] … !!!). This creates a 2(valence: positive, negative; between) x 2(arousal: low, high; within) x 3(intensity: low, medium, high; between) mixed design. The specific emotions used are in Table 2.A1 and we used the same book reviews as Study 4A. Participants still saw two book reviews, one low-arousal and one high-arousal, and they both had the same valence and intensity.

Table 2.A1: Study S1 Emotional States

Valence	**Arousal**	**Emotions**
Negative	Low	disappointed, upset, bored, dissatisfied
	High	angry, frustrated, annoyed, irritated
Positive	Low	content, pleased, satisfied, relieved
	High	happy, thrilled, excited, delighted

Results and Discussion

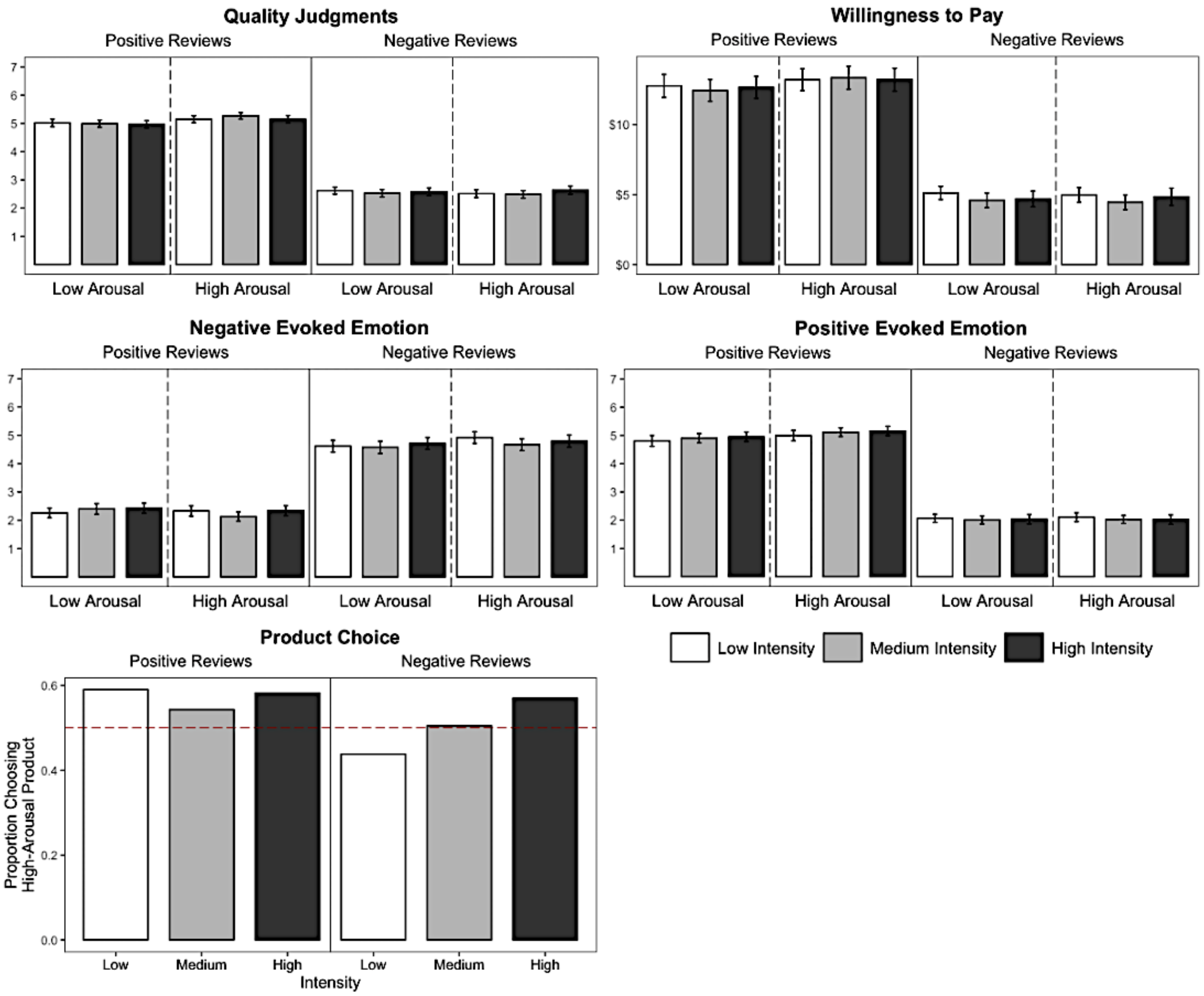
Quality Judgments and WTP. First, we regressed quality judgments on dummy variables for review valence, arousal, intensity, and all interactions (standard errors clustered by participant). We then ran the same regression for WTP. Results from these regressions are in

Table 2.A2, but they generally do not support our intensity hypotheses. Collapsing across intensity, we replicated our prior finding that high-arousal products were preferred for positive reviews (M_{High} = 5.19, SD_{High} = 0.87 vs. M_{Low} = 5.00, SD_{Low} = 0.95; t(1198) = 4.24, p < .001). However, we did not find a preference for products with low-arousal negative reviews (M_{High} = 2.54, SD_{High} = 0.98 vs. M_{Low} = 2.57, SD_{Low} = 0.92; t(1198) = -0.76, p = .445). Also, intensity did not moderate the effect. For WTP, collapsing across intensity, we similarly found arousal had a significant effect for positive reviews (M_{High} = 13.26, SD_{High} = 5.66 vs. M_{Low} = 12.62, SD_{Low} = 5.78; t(1162) = 4.14, p < .001), but not for negative reviews (M_{High} = 2.54, SD_{High} = 0.98 vs. M_{Low} = 2.57, SD_{Low} = 0.92; t(1162) = -0.45, p = .649). Again, intensity did not play a significant role.

Evoked Emotion. We ran the same regression with positive and negative evoked emotion (see Table 2.A2). For positive reviews, collapsing across intensity, positive evoked emotion was higher for high-arousal reviews than low-arousal reviews, as in prior studies (M_{High} = 5.09, SD_{High} = 1.21 vs. M_{Low} = 4.89, SD_{Low} = 1.28; t(1198) = 3.92, p < .001). For negative reviews, collapsing across intensity, negative evoked emotion was higher for high-arousal reviews than low-arousal reviews, as in prior studies (M_{High} = 4.80, SD_{High} = 1.25 vs. M_{Low} = 4.64, SD_{Low} = 1.27; t(1198) = 3.04, p = .002). However, intensity did not moderate these effects.

Product Choice. For the low-intensity condition (which essentially replicates Studies 3 and 4), we replicated our prior finding that the proportion of participants choosing the high-arousal product was significantly higher for positive reviews (59%) than negative reviews (44%; χ2 = 8.70, p = .003, Cohen's w = 0.15). However, we found no such difference when intensity was moderate or high. This does fit with our hypothesis that emotion will have less of an impact on the reader when intensity increases, but it does not map onto the patterns we see for evoked emotion.

Figure 2.A1: Study S1 Results



NOTE - All error bars are 95% confidence intervals (standard errors clustered by participant)

Table 2.A2: Study S1 Regression Results

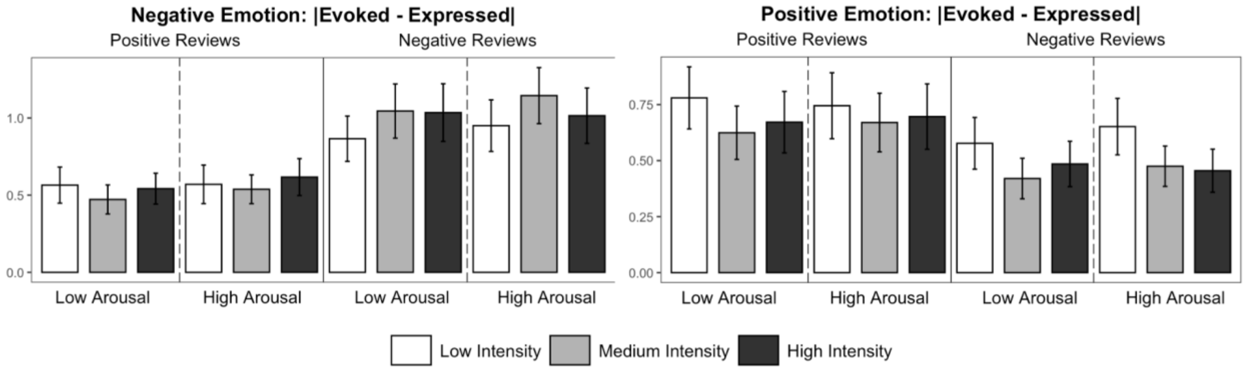
	Quality	**WTP**	**Positive Evoked**	**Negative Evoked**
Predictors	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)
(Intercept)	2.62 *** (0.06)	5.12 *** (0.24)	2.06 *** (0.07)	4.62 *** (0.11)
Valence [positive]	2.40 *** (0.09)	7.65 *** (0.48)	2.74 *** (0.12)	-2.36 *** (0.14)
Arousal [high]	-0.10 (0.06)	-0.14 (0.17)	0.04 (0.09)	0.30 ** (0.10)
Intensity [medium]	-0.09 (0.09)	-0.53 (0.36)	-0.06 (0.10)	-0.04 (0.15)
Intensity [high]	-0.04 (0.09)	-0.43 (0.38)	-0.03 (0.11)	0.10 (0.15)
Valence [positive] × Arousal [high]	0.23 * (0.10)	0.58 (0.33)	0.15 (0.13)	-0.23 (0.14)
Valence [positive] × Intensity [medium]	0.07 (0.13)	0.20 (0.68)	0.16 (0.17)	0.18 (0.20)
Valence [positive] × Intensity [high]	-0.01 (0.13)	0.32 (0.69)	0.18 (0.17)	0.07 (0.20)
Arousal [high] × Intensity [medium]	0.06 (0.09)	0.00 (0.23)	-0.02 (0.11)	-0.21 (0.13)
Arousal [high] × Intensity [high]	0.16 (0.09)	0.29 (0.23)	-0.05 (0.11)	-0.22 (0.13)
Valence [positive] × Arousal [high] × Intensity [medium]	0.08 (0.14)	0.47 (0.44)	0.04 (0.17)	-0.13 (0.19)
Valence [positive] × Arousal [high] × Intensity [high]	-0.12 (0.15)	-0.18 (0.45)	0.07 (0.17)	0.06 (0.19)
Observations	2398	2326	2398	2398
N$_{id}$	1199	1163	1199	1199

R^2	0.65	0.42	0.61	0.43
R^2 adjusted	0.65	0.42	0.61	0.42

NOTE - * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Evoked and Expressed Divergence. For each participant, we calculated an absolute difference score for the difference between evoked and expressed emotion (separately for positive and negative emotion). We then ran the same regression with these absolute difference scores as dependent variables (see Figure 2.A2 and Table 2.A3). The results are noisy but suggest an interesting asymmetry between positive and negative emotion. For positive emotion, evoked and expressed emotion were least similar when intensity was low, and converged more when intensity was moderate or high (with no difference between the moderate and high conditions). For negative emotion, evoked and expressed emotion were most similar when intensity was low, and diverged slightly more when intensity was moderate or high (with no difference between the moderate and high conditions).

Figure 2.A2: Study S1 |Evoked Emotion - Expressed Emotion|



NOTE - All error bars are 95% confidence intervals (standard errors clustered by participant)

Table 2.A3: Study S1 |Evoked Emotion - Expressed Emotion|

	| Evoked - Expressed | Positive	**| Evoked - Expressed | Negative**
Predictors	Estimates (S.E.)	Estimates (S.E.)
(Intercept)	0.58 *** (0.06)	0.87 *** (0.07)
Valence [positive]	0.20 * (0.09)	-0.30 ** (0.10)
Arousal [high]	0.07 (0.08)	0.08 (0.09)
Intensity [medium]	-0.16 * (0.07)	0.18 (0.12)
Intensity [high]	-0.09 (0.08)	0.17 (0.12)
Valence [positive] × Arousal [high]	-0.11 (0.10)	-0.08 (0.12)
Valence [positive] × Intensity [medium]	0.00 (0.12)	-0.27 (0.14)
Valence [positive] × Intensity [high]	-0.02 (0.13)	-0.19 (0.14)
Arousal [high] × Intensity [medium]	-0.02 (0.10)	0.02 (0.12)
Arousal [high] × Intensity [high]	-0.10 (0.10)	-0.10 (0.13)
Valence [positive] × Arousal [high] × Intensity [medium]	0.10 (0.13)	0.05 (0.16)
Valence [positive] × Arousal [high] × Intensity [high]	0.16 (0.13)	0.17 (0.16)
Observations	2398	2398

N_{id}	1999	1999
R^2	0.02	0.05
R^2 adjusted	0.01	0.05

NOTE - * $p<0.05$ ** $p<0.01$ *** $p<0.001$

*Discussion***.** Study S1 investigated the role of emotional intensity as a potential moderator of the joint impact of valence and arousal on evoked emotion and product evaluations. We replicated some of our findings from previous studies and did find evidence that increasing intensity attenuated the effect of valence on the propensity to choose the high-arousal product. However, in our other key measures, we generally do not find evidence in support of intensity as a moderator. Future work could explore why this is the case or attempt to manipulate intensity in other ways to further investigate the role it plays in consumer behavior.

The other goal of Study S1 was to test the effect of intensity on the absolute difference between evoked and expressed emotion. We hypothesized intensity could be a key factor in that relationship, predicting the two would diverge more as intensity increases. While we found suggestive evidence of that for negative emotion, we observed the opposite for positive emotions. Future research could further explore this asymmetry to study whether and/or why the relationship between evoked and expressed emotion differs by valence.

Appendix 2.B: Full Regression Tables and Mediation Results

Study 1

Full Regression Results

	Expressed Emotion Only	Evoked Emotion Only	Evoked & Expressed Emotion
Predictors	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)
(Intercept)	2.88 (1.61)	3.15 * (1.50)	3.85 ** (1.48)
Negative Expressed Emotion	-0.24 *** (0.03)	-0.03 (0.04)	
Positive Expressed Emotion	0.47 *** (0.03)	0.26 *** (0.04)	
Negative Evoked Emotion		-0.30 *** (0.04)	-0.40 *** (0.03)
Positive Evoked Emotion		0.26 *** (0.04)	0.45 *** (0.03)
Helpfulness	-0.05 (0.05)	-0.05 (0.05)	-0.06 (0.05)
Objectivity	0.04 (0.03)	0.04 (0.03)	0.04 (0.03)
Vividness	0.11 ** (0.04)	0.11 ** (0.04)	0.10 ** (0.04)
Informativeness	0.03 (0.05)	0.02 (0.04)	0.03 (0.04)
Reviewer Reliability	-0.02 (0.05)	-0.03 (0.05)	-0.03 (0.05)
Perceived Similarity to Reviewer	0.06 (0.04)	0.04 (0.04)	0.03 (0.04)

Perceived Reviewer Experience	0.00 (0.02)	0.02 (0.02)	0.01 (0.02)
PANAS Negative	-0.03 * (0.01)	-0.02 (0.01)	-0.01 (0.01)
PANAS Positive	-0.00 (0.01)	0.00 (0.01)	0.01 (0.01)
Number of Reviews	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Average Rating	0.09 (0.35)	0.04 (0.33)	-0.04 (0.33)
Random Effects			
τ_{00}	0.03 id	0.05 id	0.04 id
	0.12 product	0.10 product	0.10 product
N	601 id	601 id	601 id
	24 product	24 product	24 product
Observations	1202	1202	1202
Marginal R^2	0.45	0.50	0.48
Conditional R^2	0.52	0.57	0.55
AIC	3680.34	3569.38	3612.07

NOTE - * p<0.05 ** p<0.01 *** p<0.001

Study 2

Full Regression Results

	Quality Judgments	**WTP**	**Negative Evoked**	**Positive Evoked**	**Bipolar Evoked**
Predictors	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)
(Intercept)	3.91 *** (0.04)	7.31 *** (0.19)	3.65 *** (0.09)	3.54 *** (0.09)	3.82 *** (0.05)

Review type [emotional]	0.08 (0.05)	0.39 ** (0.12)	0.06 (0.09)	0.36 *** (0.09)	0.06 (0.07)
Arousal [high]	-0.03 (0.06)	0.07 (0.27)	-0.08 (0.12)	0.07 (0.13)	0.07 (0.08)
Review type [emotional] × Arousal [high]	0.06 (0.07)	0.09 (0.18)	0.21 (0.14)	0.04 (0.13)	-0.08 (0.10)
Observations	1994	1802	994	994	1000
R^2	0.00	0.00	0.01	0.02	0.00
R^2 adjusted	0.00	0.00	0.00	0.02	-0.00

NOTE - * p<0.05 ** p<0.01 *** p<0.001

Study 3

Full Regression Results

	Quality Judgments	WTP	Negative Evoked	Positive Evoked
Predictors	Estimates (S.E)	Estimates (S.E)	Estimates (S.E.)	Estimates (S.E.)
(Intercept)	3.11 *** (0.11)	-0.23 ** (0.09)	1.98 *** (0.11)	0.46 *** (0.09)
Valence [positive]	2.30 *** (0.05)	0.43 *** (0.03)	-1.52 *** (0.04)	1.49 *** (0.04)
Arousal [high]	-0.31 *** (0.05)	-0.06 † (0.03)	0.54 *** (0.04)	-0.12 ** (0.04)
Valence [positive] × Arousal [high]	0.66 *** (0.07)	0.14 ** (0.05)	-0.61 *** (0.06)	0.87 *** (0.06)
Helpfulness	-0.01 (0.02)	-0.02 (0.02)	0.11 *** (0.02)	0.22 *** (0.02)
Objectivity	0.01 (0.02)	0.00 (0.01)	0.01 (0.02)	0.01 (0.02)

Disregarding of Information	-0.01 (0.02)	0.04^{**} (0.01)	0.22^{***} (0.02)	0.06^{***} (0.02)
Random Effects				
σ^2	0.85	0.37	0.60	0.62
τ_{00}	0.14_{id}	0.56_{id}	0.10_{id}	0.15_{id}
	0.01_{stim}	0.00_{stim}	0.02_{stim}	0.00_{stim}
ICC	0.15	0.60	0.17	0.20
N	728_{id}	728_{id}	728_{id}	728_{id}
	4_{stim}	4_{stim}	4_{stim}	4_{stim}
Observations	2912	2912	2912	2912
Marginal R^2	0.64	0.06	0.59	0.57
Conditional R^2	0.69	0.63	0.66	0.66

NOTE - † $p<0.1$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Moderated Mediation Results (Quality Judgments)

Effect	Arousal	Estimate	95% Bootstrapped CI
Indirect via Positive Evoked	Low	0.69	[0.62, 0.77]
	High	1.09	[0.99, 1.20]
Indirect via Negative Evoked	Low	0.52	[0.45, 0.60]
	High	0.73	[0.63, 0.84]
Direct	Low	1.10	[0.98, 1.21]
	High	1.14	[1.00, 1.28]

Moderated Mediation Results (WTP)

Effect	Arousal	Estimate	95% Bootstrapped CI
Indirect via Positive Evoked	Low	0.28	[0.19, 0.37]
	High	0.44	[0.30, 0.58]
Indirect via Negative Evoked	Low	-0.04	[-0.11, 0.03]
	High	-0.06	[-0.16, 0.04]
Direct	Low	0.25	[0.12, 0.37]
	High	0.26	[0.11, 0.42]

Study 4A

Full Regression Results

	Quality Judgments	**WTP**	**Negative Evoked**	**Positive Evoked**
Predictors	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)
(Intercept)	2.61^{***} (0.07)	6.28^{***} (0.35)	3.83^{***} (0.11)	2.24^{***} (0.09)
Valence [positive]	2.41^{***} (0.09)	6.25^{***} (0.59)	-1.59^{***} (0.15)	2.23^{***} (0.14)
Arousal [high]	$-0.12^{\dagger}$ (0.07)	-0.78^{**} (0.24)	0.32^{***} (0.09)	-0.18^{*} (0.08)
Valence [positive] × Arousal [high]	0.30^{**} (0.10)	1.60^{***} (0.40)	-0.51^{***} (0.14)	0.45^{***} (0.13)
Observations	698	686	698	698
R^2	0.68	0.29	0.30	0.47
R^2 adjusted	0.68	0.29	0.30	0.46

NOTE - † $p<0.1$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Moderated Mediation Results (Quality Judgments)

Effect	**Arousal**	**Estimate**	**95% Bootstrapped CI**
Indirect via Positive Evoked	Low	0.49	[0.36, 0.64]
	High	0.59	[0.44, 0.74]
Indirect via Negative Evoked	Low	0.17	[0.09, 0.25]
	High	0.22	[0.12, 0.33]
Direct	Low	1.75	[1.54, 1.96]
	High	1.91	[1.68, 2.13]

Moderated Mediation Results (WTP)

Effect	Arousal	Estimate	95% Bootstrapped CI
Indirect via Positive Evoked	Low	1.21	[0.48, 1.96]
	High	1.47	[0.60, 2.36]
Indirect via Negative Evoked	Low	0.58	[0.12, 1.09]
	High	0.77	[0.16, 1.41]
Direct	Low	4.46	[3.06, 5.86]
	High	5.61	[4.09, 7.13]

Study 4B

Full Regression Results

	Quality Judgments	WTP	Negative Evoked	Positive Evoked
Predictors	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)
(Intercept)	2.34 *** (0.07)	17.64 *** (1.10)	4.72 *** (0.08)	1.82 *** (0.09)
Valence [positive]	2.96 *** (0.10)	32.21 *** (1.56)	-3.01 *** (0.12)	3.01 *** (0.13)
Arousal [high]	-0.09 (0.09)	-0.30 (0.91)	0.21 * (0.08)	-0.04 (0.11)
Valence [positive] × Arousal [high]	0.27 (0.12)	2.98 (1.28)	-0.34 * (0.12)	0.41 * (0.15)
Random Effects				
σ^2	0.43	60.97	0.59	0.62
τ_{00}	0.44_{id}	322.28_{id}	1.44_{id}	0.84_{id}
	$0.01_{emotion}$	$1.04_{emotion}$	$0.01_{emotion}$	$0.02_{emotion}$
ICC	0.51	0.84	0.71	0.58
N	$16_{emotion}$	$16_{emotion}$	$16_{emotion}$	$16_{emotion}$
	800_{id}	800_{id}	800_{id}	800_{id}
Observations	1600	1600	1600	1600

Marginal R^2	0.73	0.43	0.56	0.64
Conditional R^2	0.87	0.91	0.87	0.85

NOTE - * p<0.05   ** p<0.01   *** p<0.001

Moderated Mediation Results (Quality Judgments)

Effect	Arousal	Estimate	95% Bootstrapped CI
Indirect via Positive Evoked	Low	0.62	[0.49, 0.75]
	High	0.71	[0.57, 0.85]
Indirect via Negative Evoked	Low	0.31	[0.21, 0.42]
	High	0.35	[0.24, 0.47]
Direct	Low	2.02	[1.84, 2.20]
	High	2.17	[1.98, 2.37]

Moderated Mediation Results (WTP)

Effect	Arousal	Estimate	95% Bootstrapped CI
Indirect via Positive Evoked	Low	7.70	[5.04, 10.45]
	High	8.76	[5.79, 11.92]
Indirect via Negative Evoked	Low	2.17	[0.19, 4.20]
	High	2.42	[0.21, 4.65]
Direct	Low	22.34	[18.41, 26.26]
	High	24.02	[19.82, 28.22]

Study 5

Full Regression Results

	Quality Judgments	**WTP**	**Negative Evoked**	**Positive Evoked**
Predictors	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)	Estimates (S.E.)
(Intercept)	3.21 *** (0.20)	4.92 *** (0.89)	2.28 *** (0.31)	1.99 *** (0.26)

Valence [positive]	0.90 *** (0.20)	2.62 ** (0.99)	0.23 (0.31)	0.06 (0.25)
Arousal	-0.03 (0.03)	-0.20 (0.12)	0.36 *** (0.05)	-0.11 ** (0.04)
Valence [positive] × Arousal	0.29 *** (0.04)	0.83 *** (0.20)	-0.51 *** (0.06)	0.52 *** (0.05)
Helpfulness	-0.05 (0.03)	-0.00 (0.16)	0.01 (0.04)	0.08 * (0.04)
Objectivity	-0.03 (0.02)	-0.14 (0.11)	0.02 (0.03)	0.01 (0.02)
Reliability	0.02 (0.03)	0.18 (0.17)	0.04 (0.05)	0.08 * (0.04)
Informativeness	-0.02 (0.03)	0.10 (0.17)	0.08 * (0.04)	-0.01 (0.03)
Observations	3196	2956	3196	3196
R^2	0.54	0.32	0.40	0.57
R^2 adjusted	0.54	0.32	0.40	0.57

NOTE - * p<0.05   ** p<0.01   *** p<0.001

Moderated Mediation Results (Quality Judgments)

Effect	Arousal Rating	Estimate	95% Bootstrapped CI
Indirect via Positive Evoked	16th Percentile	0.47	[0.40, 0.52]
	50th Percentile	0.64	[0.55, 0.73]
	84th Percentile	0.76	[0.65, 0.87]
Indirect via Negative Evoked	16th Percentile	0.16	[0.11, 0.21]
	50th Percentile	0.23	[0.16, 0.30]
	84th Percentile	0.28	[0.19, 0.36]
Direct	16th Percentile	1.35	[1.23, 1.46]
	50th Percentile	1.51	[1.39, 1.62]
	84th Percentile	1.62	[1.48, 1.77]

Moderated Mediation Results (WTP)

Effect	Arousal Rating	Estimate	95% Bootstrapped CI
Indirect via Positive Evoked	16th Percentile	1.41	[1.08, 1.76]
	50th Percentile	1.78	[1.36, 2.19]
	84th Percentile	2.26	[1.75, 2.80]
Indirect via Negative Evoked	16th Percentile	0.34	[0.12, 0.56]
	50th Percentile	0.46	[0.16, 0.74]
	84th Percentile	0.62	[0.22, 1.00]
Direct	16th Percentile	3.90	[3.33, 4.47]
	50th Percentile	4.25	[3.69, 4.80]
	84th Percentile	4.71	[3.96, 5.47]

Appendix 2.C: Study Stimuli

Study 1

Product Category	Products Used
Appliances	Ice maker, washing machine, freezer, dishwasher
Home and kitchen	Waffle maker, mattress protector, pot, coffee maker
Electronics	Camera, speaker, watch, TV
Sports and fitness	Exercise bike, elliptical, resistance bands, portable home workout
Tools and home improvement	Drill, hose, shower head, fan
Entertainment	One movie and three different TV shows

NOTE - For full text of the top positive and negative review for each product, see: https://osf.io/wrbhf/?view_only=8a4603b1bb4548acbb0ff373d75576be

Study 2

Review Type	Review Text
Bland	Average book. Every chapter ended on a suspenseful note that made me want to keep reading and the author's use of imagery was great. However, the ending wasn't great and a lot of the dialogue felt forced.
Emotional (Low Arousal)	Average book. I was pleased with how interesting the plot was and grateful the author did a good job developing the characters. However, I was disappointed because the author spent too much time in the beginning setting the scene and upset that the writing was confusing at times.
Emotional (High Arousal)	Average book. I was thrilled with how interesting the plot was and I enjoyed the way the author developed the characters. However, I was irritated that the author spent too much time in the beginning setting the scene and hated that the writing was confusing at times.

NOTE – The product information (i.e., the text other than the emotion words) was randomly assigned to the bland or emotional review. Participants saw only one emotional review. For full study text, see: https://osf.io/wrbhf/?view_only=8a4603b1bb4548acbb0ff373d75576be

Study 3

Product	Valence	Arousal	Review Text
Exercise Bike	Negative	Low	**Review Title**: Pedals not very durable **Review**: The bike was fairly easy to set up and it folded up into a pretty compact size. My wife and I each rode it about three times a week. After about a month, she was riding it one morning and the left pedal popped off. I tried reattaching the pedal, but the plastic had snapped. The pedals don't seem to be quite as sturdy as advertised.
		High	**Review Title**: Issues with bike seat. Unbelievable. Very upset! **Review**: The assembly for this bike was relatively straightforward, but I noticed that the seat is at a bit of a weird angle and is somewhat misaligned with the frame. Once I tried it out, I realized that this results in an awkward riding position. I really hate this riding position. The seat also wobbles even when the knob is fully

			tightened, so I don’t enjoy riding the bike. I followed the instructions very closely and reviewed them after I encountered this issue, but nothing changed, so that was useless. I am so upset. I was feeling so good about doing something good for my body and now I’m just frustrated that I purchased it.
	Positive	Low	**Review Title**: Sturdy bike with useful pedal feature **Review**: The bike was fairly easy to set up and it folded up into a pretty compact size. For the past month, my wife and I each used it about three times a week. The bike provides a range of resistance modes and we have had no problems with it so far. It also has a useful feature that allows you to strap your foot to pedal for a more secure ride.
		High	**Review Title**: Wonderful bike! Reliable with helpful seat features. Very happy. **Review**: The assembly for this bike was relatively straightforward, which was great. I’ve been using it for about three weeks now and haven’t had any problems. I wish I could ride this bike all the time! There are a range of different resistance modes that are available. The seat has a tilt feature that allows you to make the seat level will in a recumbent position. The seat also has a back pad that provides a more comfortable ride. I love these features! I am so happy. It feels great to do something good for body and I’m really grateful that I purchased it.
Toaster	Negative	Low	**Review Title:** Takes a long time to cook **Review**: I tested out this toaster oven by baking chicken parmesan for the family on Friday night. I set the oven to 350 degrees, waited 15 minutes for it to heat up, and put the chicken in. I pulled the chicken out 35 minutes later. When we cut into the chicken, it was still raw inside. The reheating function of this oven might work okay, but I don’t think that this product cooks meat very well.
		High	**Review Title**: Not very effective at baking bread. Incredibly frustrated! **Review**: I was really excited about this toaster oven, but

			it totally disappointed. The morning after this toaster oven arrived, I tried using it to bake a loaf of bread. When I checked on the bread, about halfway through the cooking time, the loaf was already burnt. I couldn’t believe it. I tried again, this time with the oven on a slightly lower setting. Once the baking time was up, I took the loaf out to let it rest for about 90 minutes. When I cut into the loaf, it was still raw in the middle. So disgusting. I really don’t have time for this. I thought this oven would be so handy to have, but now I’m just very upset that I purchased it.
	Positive	Low	**Review Title**: Cooks meat effectively and works as a microwave **Review**: I tested out this toaster oven by baking chicken parmesan for the family on Friday night. I set the oven to 350 degrees, waited 15 minutes for it to heat up, and put the chicken in. I pulled the chicken out 35 minutes later. When we cut into the chicken, it was cooked properly -- all the way through. It seems like this oven functions effectively as both a microwave and an oven.
		High	**Review Title**: So impressed! Effective at baking bread without needing much attention. Excellent oven. **Review:** I was really excited about this toaster oven and it did not disappoint! The morning after this toaster oven arrived, I tried using it to bake a loaf of bread. After preparing the dough, I set a timer for about half the time, and left the kitchen to do some work. When I came back to check on the bread, it seemed to be baking properly. Once the baking time was up, I took the loaf out to let it rest for about 90 minutes. When I bit into the loaf it was perfection, cooked beautifully -- all the way through. Such delicious bread. I’ve now tried baking several different kinds of bread and it works like a charm each time. I really love this appliance and can’t wait to do more cooking with it!

NOTE – The product information (i.e., the text other than the emotion words) was randomly assigned to the low- and high-arousal conditions. For full study text, see: https://osf.io/wrbhf/?view_only=8a4603b1bb4548acbb0ff373d75576be

Study 4A

Valence	Review Text
Negative	This book wasn't great. The plot was very boring, not much happened. The characters were not well developed and a lot of the dialogue seemed forced. The author's writing was confusing at times but the end wraps things up clearly.
	Overall, it was pretty underwhelming. There was one point where I thought the story was going somewhere but then it went back to mundane. The writing in general could have been better, some parts get very repetitive.
Positive	Good book that captivated me from the onset. The author does a great job of really making you care about the characters. Ending wasn't spectacular but I would definitely recommend it.
	This book was hard to put down, every chapter ended on a suspenseful note to make the reader continue. The plot twists and turns kept things interesting but were overdone a bit. Overall, it was a great book.

NOTE – The headline of the review was randomized (to manipulate arousal)

Study 4B

Valence	Reviews Used
Negative	This blender wasn't great. It is not powerful enough to blend tough items and makes too much noise. I would definitely not recommend it.
	I did not like this blender. There were only a few settings that weren't very helpful and it was extremely difficult to clean. Overall, this was a bad purchase.
Positive	This is a great blender. It is powerful enough to blend almost anything and doesn't make too much noise. I would definitely recommend it.
	I really like this blender. It has plenty of helpful modes and settings and is extremely easy to clean. Overall, this was a great purchase.

NOTE – The headline of the review was randomized (to manipulate arousal)

Study 5

Emotion	Review Text
Surprise	This book is an exhilarating mix of suspense and drama. The plot is full of unexpected twists that keep readers hooked. Characters are intriguing and unique, enhancing the narrative. While some moments are bizarre and bewildering, the overall experience is wonderfully immersive. The abrupt ending leaves some curiosity unfulfilled, but the journey is definitely worth it.
	This book is a thrilling mix of suspense and drama. The plot, filled with abductions and catastrophes, keeps readers engaged and on edge. The author did an outstanding job developing spectacular and intiguing characters. However, the narrative sometimes feels a bit confusing. Despite this, the story remains engaging and immersive.
	This book is an entertaining journey. The protagonist faces abrupt catastrophes, from abduction to ambush, with many captivating moments. The urgent pace and constant twists keep the reader in astonishment, though the unpredictability can be overwhelming. Some plot points feel erratic, leaving readers occasionally bewildered. Overall, a captivating read with a dynamic plot and memorable characters.
Anticipation	This was an exciting and suspenseful book. The protagonist's journey captivates, balancing peril and thrill perfectly. Characters are both relatable and admirable, adding depth to the story. Some parts feel overly ambitious, but the overall experience remained exciting. The plot keeps you engaged with a good mix of tension and anticipation. Despite some minor flaws, it’s a rewarding read.
	This was a great book filled with suspense and excitement. Characters are relatable and admirable, adding depth to the story. Although some parts feel overly ambitious, the overall experience is thrilling and enjoyable. The plot keeps you engaged with tension and anticipation, making it hard to put down. Despite some clear flaws, it’s a spectacular book.
	The book provides a thrilling adventure. The characters have ambition and curiosity, yet the plot feels a bit drawn out. There is excitement and some suspense, with a few twists that captivate. However, the pacing can seem slow, and the climax doesn't entirely live up to the anticipation. Despite some flaws, the overall experience is exciting and engaging.
Joy	An accomplished and beautifully written story, this book is a delightful read. The characters are wonderfully alive and the plot, albeit predictable, is

	engaging. The author's ability to create vivid and aesthetic scenes is admirable. However, despite the abundant charm and allure, the pacing is slow at times. Nevertheless, the heartfelt emotions and rich detail make it a rewarding experience.
	The book is a delightful read. The characters are lovable, and the story, though not without flaws, is engaging and amusing. The author's ability to create a beautiful, vibrant world is admirable. However, at times, the plot feels overly ambitious and the pacing can be a bit uneven. Nevertheless, the overall experience is enjoyable, making it an outstanding addition to your collection.
	This book offers an enjoyable narrative with many charming and memorable moments. The characters are well-developed, and their journey is filled with ambition and affection. While the plot has some predictable elements, the author's writing style is delightful and engaging. Overall, it is a lovely read that leaves a lasting impression, though it could use a bit more excitement.
Trust	This was a great book. The author’s understanding of human nature is evident and admirable. The characters are credible, and the narrative is engaging. The author's ability to convey emotions is impressive. However, the plot lacks complexity. The pace could be quicker, but the themes of loyalty and love are well-explored. Overall, it is an enjoyable read with some memorable moments.
	This book is commendable. The author provides an accurate and engaging narrative. The protagonist showcases unwavering determination, integrity, and loyalty. I love the plot but the story is somewhat predictable. The supporting characters, while likable, lack depth. Overall, this book deserves praise for its captivating storytelling and admirable themes. It is a worthwhile read.
	This book offers a blend of charming narrative and admirable storytelling. The author’s intelligence and expertise is evident. The themes of loyalty and trust resonate throughout. However, the plot occasionally lacks depth, and some characters, while lovable, feel underdeveloped. Despite these minor shortcomings, the novel’s overall merit and warmhearted tone make it a delightful read.
Disgust	The book was an awkward read. The plot was abominable, filled with annoying actions and frustrating characters. It felt like a grotesque mix of bad ideas, leading to a grim experience. Dialogue was mostly bickering and

	insults, which was off-putting. Despite a few redeeming moments, the overall experience was disappointing.
	This book definitely had potential but ended up being almost unbearable. The plot is a travesty with annoying characters and atrocious dialogue. The protagonist is obnoxious, and the story is bogged down with unnecessary, awkward details. Despite a few interesting moments, the overall experience is disappointing.
	The book had potential but was ultimately disappointing. The story was plagued by abhorrent events and disgusting characters. Despite moments of genuine interest, the overwhelming sense of alienation and animosity among the characters makes it a difficult read. The abundance of adverse elements detracts from any potential enjoyment and left me unsatisfied.
Anger	This book is marred by confusion and constant conflict. The characters are often angry and alienated, entangled in endless arguments and antagonism. While there are occasional good scenes, the overall narrative is filled with aggression and despair. It’s a bitter, tumultuous read that falls short of being enjoyable or engaging.
	This book, despite some interesting moments, ultimately left me disappointed. The story felt chaotic and often confusing. The characters were plagued by adversity, constantly facing conflict and hardship. The narrative was filled with aggressive confrontations and antagonistic interactions. Although some parts had potential, the overall execution fell short, leading to an unsatisfying experience.
	This book had a promising start but quickly descended into a mess of conflict and confusion. The plot was hindered by relentless adversity and antagonistic characters. There were some moments of potential, but the story was marred by incessant aggression and chaotic interactions. Ultimately, the narrative failed to engage, leaving me feeling disappointed and disheartened.
Fear	The book had a lot of potential but ended up being somewhat disappointing. The storyline was filled with anxiety and confusion. The characters often faced distressing situations and the overall atmosphere was dark and gloomy. There were moments of tension and alarm that felt overly dramatic. Despite a few redeeming qualities, the constant sense of dread and turmoil overshadowed any positive aspects.
	The book had an intriguing premise but ultimately fell short. The plot was fraught with confusion and anxiety, making it hard to follow. The characters seemed to be constantly in distress. There were some suspenseful

	moments, but they often felt forced and overly dramatic. Despite a few interesting ideas, the story was overshadowed by a sense of doom and despair.
	This book had an interesting start but quickly became disappointing. The plot was filled with confusion and anxiety, making it hard to stay engaged. The characters often faced distressing and alarming situations, which felt overdone. The overall tone was dark and filled with a sense of dread, which overshadowed the few moments of excitement and led to an ultimately unsatisfying read.
Sadness	The book had its moments, but overall, it felt a bit lacking. The story seemed aimless at times, and some parts were uninteresting. The characters were not as engaging as they could be, leading to a sense of detachment. The plot had potential but felt unresolved by the end. There were some interesting ideas and moments that showed promise, but not enough to recommend.
	The book had a lot of potential but fell short in many aspects. The story was often confusing and hard to follow. The characters felt flat and lacked depth. The plot had some interesting moments, but they were overshadowed by too much sadness and despair. Despite a few redeeming qualities, but the book left me feeling disappointed and unfulfilled.
	The book was quite disappointing. Despite the potential, it was plagued by an incoherent plot and poorly developed characters. The story felt dragged out and ultimately went nowhere, leaving me feeling unfulfilled. The ending was abrupt and unsatisfying, lacking any real closure. I had hoped for more, but the book left me feeling let down and disinterested.

NOTE – Participants saw one of the three reviews for each emotion they were assigned to

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.

Bao, Z., & Chau, M. (2016, June). The effect of Collective Rating on the perception of Online Reviews. In *PACIS* (p. 123).

Barrett, L. F., & Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current directions in psychological science*. *8*(1), 10-14.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bauer, R. A. (1960). Consumer behavior as risk. *American Marketing Association*, 389–398.

Bechara, A., Damasio, H., Tranel, D., and Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, *275*(5304), 1293–1295. https://doi.org/10.1126/science.275.5304.1293

Berger, J., & Milkman, K. L. (2016). What Makes Online Content Viral ? What Makes Online Content Viral ? *American Marketing Association*, *49*(December 2009), 192–205. www.marketingpower.com/jmr_

Caprariello, P. A., & Reis, H. T. (2013). To do, to have, or to share? Valuing experiences over material possessions depends on the involvement of others. *Journal of personality and social psychology*, *104*(2), 199.

Casaló, L. V., Flavián, C., Guinalíu, M., & Ekinci, Y. (2015). Avoiding the dark side of positive online consumer reviews: Enhancing reviews' usefulness for high risk-averse travelers. *Journal of Business Research*, *68*(9), 1829-1835.

Chang, Y. C., Chu, C. H., Chen, C. C., & Hsu, W. L. (2016, June). Linguistic template extraction for recognizing reader-emotion. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 21, Number 1, June 2016*.

Chen, Z., & Lurie, N. H. (2013). Temporal contiguity and negativity bias in the impact of online word of mouth. *Journal of Marketing Research*, *50*(4), 463-476.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. In *Journal of Marketing Research* (Vol. 43, Issue 3, pp. 345–354). SAGE PublicationsSage CA: Los Angeles, CA. https://doi.org/10.1509/jmkr.43.3.345

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, *29*(5), 944–957. https://doi.org/10.1287/mksc.1100.0572

Dai, H., Chan, C., & Mogilner, C. (2020). People rely Less on consumer reviews for experiential than material purchases. *Journal of Consumer Research*, *46*(6), 1052-1075.

Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., & Lee, L. (2009, April). How opinions are received by online communities: a case study on amazon. com helpfulness votes. In Proceedings of the 18th international conference on World wide web (pp. 141-150).

Dave, C., & Wolfe, K. W. (2003). On confirmation bias and deviations from Bayesian updating

de Langhe, B., Fernbach, P. M., & Lichtenstein, D. R. (2016). Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*, *42*(6), 817–833. https://doi.org/10.1093/jcr/ucv047

Diener, E. (1999). Introduction to the special section on the structure of emotion. *Journal of personality and Social Psychology*, *76*(5), 803.

Doh, S. J., & Hwang, J. S. (2009). How consumers evaluate eWOM (electronic word-of-mouth) messages. *CyberPsychology and Behavior*, *12*(2), 193-197.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, *13*(1), 1–17. https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S

Fischer, P. (2011). Selective exposure, decision uncertainty, and cognitive economy: A new theoretical perspective on confirmatory information search. *Social and Personality Psychology Compass*, *5*(10), 751-762.

Fisher, M., Newman, G. E., & Dhar, R. (2018). Seeing stars: How the binary bias distorts the interpretation of customer ratings. *Journal of Consumer Research*, *45*(3), 471-489.

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.

Galante, M. (2018). *People Think User-Generated Reviews are More Trustworthy than Professional Reviews*. Squareup. https://squareup.com/us/en/townsquare/people-think-user-generated-reviews-are-more-trustworthy

Ghose, A., & Ipeirotis, P. G. (2010). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE transactions on knowledge and data engineering*, *23*(10), 1498-1512.

Gino, F. (2015). *Don't Let Emotions Screw Up Your Decisions*. Harvard Business Review. https://hbr.org/2015/05/dont-let-emotions-screw-up-your-decisions

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Han, S., Lerner, J. S., & Keltner, D. (2000). Feelings and Consumer Decision Making. *Journal of Consumer Pschology*, *17*(3), 158–168. https://onlinelibrary.wiley.com/doi/pdf/10.1016/S1057-7408(07)70023-2

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate behavioral research*, *50*(1), 1-22.

Heck, D. W., Seiling, L., & Bröder, A. (2020). The love of large numbers revisited: A coherence model of the popularity bias. *Cognition*, *195*, 104069.

Higgins, E. T. (1987). Self-discrepancy: a theory relating self and affect. *Psychological review*, *94*(3), 319.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, *24*(1), 1-55.

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

Johnson, E. J., & Tversky, A. (1983). Affect, generalization, and the perception of risk. *Journal of Personality and Social Psychology*, *45*(1), 20–31. https://doi.org/10.1037/0022-3514.45.1.20

Kaemingk, D. (2020). *Online Review Statistics to Know in 2021* . Qualtrics. https://www.qualtrics.com/blog/online-review-stats/

Kim, J., & Gupta, P. (2012). Emotional expressions in online user reviews: How they influence consumers' product evaluations. *Journal of Business Research*, *65*(7), 985–992. https://doi.org/10.1016/j.jbusres.2011.04.013

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, *94*(2), 211.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological bulletin*, *110*(3), 499.

Kronrod, A., & Danziger, S. (2013). "Wii Will Rock You!" The Use and Effect of Figurative Language in Consumer Reviews of Hedonic and Utilitarian Consumption. *JOURNAL OF CONSUMER RESEARCH, Inc. •*, *40*. https://doi.org/10.1086/671998

Kupor, D., & Tormala, Z. (2018). When moderation fosters persuasion: The persuasive power of deviatory reviews. *Journal of Consumer Research*, *45*(3), 490-510.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest Package: Tests in Linear Mixed Effects Models . *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*, *81*(1), 146–159. https://doi.org/10.1037/0022-3514.81.1.146

Lerner, J. S., Gonzalez, R. M., Small, D. A., & Fischhoff, B. (2003). Effects of fear and anger on perceived risks of terrorism: A national field experiment. *Psychological Science*, *14*(2), 144–150. https://doi.org/10.1111/1467-9280.01433

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology*, *66*(1), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Loewenstein, G. F., Hsee, C. K., Weber, E. U., Welch, N., Hsee, C. K., Welch, N., Weber, E. U., & Welch, N. (2001). Risk as Feelings. *Psychological Bulletin*, *127*(2), 267–286. https://doi.org/10.1037/0033-2909.127.2.267

Loureiro, F., Garcia-Marques, T., & Wegener, D. T. (2020). Norms for 150 consumer products: Perceived complexity, quality objectivity, material/experiential nature, perceived price, familiarity and attitude. *PloS one*, *15*(9), e0238848.

Ludwig, S., de Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates. *Journal of Marketing*, *77*(1), 87–103. https://doi.org/10.1509/jm.11.0560

Medin, D. L., Ross, B. H., & Markman, A. B. (2005). *Cognitive psychology* (4th ed.). John Wiley & Sons.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement* (pp. 201-237). Woodhead Publishing.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175-220.

Pan, Y., & Zhang, J. Q. (2011). Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews. *Journal of Retailing*, *87*(4), 598–612. https://doi.org/10.1016/J.JRETAI.2011.05.002

Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, *90*(1), 175.

Peysakhovich, A., & Karmarkar, U. R. (2016). Asymmetric effects of favorable and unfavorable information on decision making under ambiguity. *Management Science*, *62*(8), 2163-2178.

Pieters, R. (2017). Meaningful mediation analysis: Plausible causal inference and informative communication. *Journal of Consumer Research*, *44*(3), 692–716. https://doi.org/10.1093/jcr/ucx081

Podium. (2017). *Online Review Stats: Podium State of Online Reviews* . https://www.podium.com/resources/podium-state-of-online-reviews/

Powell, D., Yu, J., DeWolf, M., & Holyoak, K. J. (2017). The love of large numbers: A popularity bias in consumer choice. *Psychological science*, *28*(10), 1432-1442.

PowerReviews. (2016). *The Power of Reviews*. https://www.powerreviews.com/wp-content/uploads/2016/04/PowerofReviews_2016.pdf

Qiu, L., Pang, J., & Lim, K. H. (2012). Effects of conflicting aggregated rating on eWOM review credibility and diagnosticity: The moderating role of review valence. Decision Support Systems, 54(1), 631-643.

Quaschning, S., Pandelaere, M., & Vermeir, I. (2015). When Consistency Matters: The Effect of Valence Consistency on Review Helpfulness. *Journal of Computer-Mediated Communication*, *20*(2), 136–152. https://doi.org/10.1111/JCC4.12106

Reisenzein, R. (1994). Pleasure-arousal theory and the intensity of emotions. *Journal of personality and social psychology*, *67*(3), 525.

Reynolds, K. E., Folse, J. A. G., & Jones, M. A. (2006). Search regret: Antecedents and consequences. *Journal of Retailing*, *82*(4), 339-348.

Rocklage, M. D., & Fazio, R. H. (2020). The Enhancing Versus Backfiring Effects of Positive Emotion in Consumer Reviews. *Journal of Marketing Research*, *57*(2), 332–352. https://doi.org/10.1177/0022243719892594

Rocklage, M. D., Rucker, D. D., & Nordgren, L. F. (2021). Mass-scale emotionality reveals human behaviour and marketplace success. *Nature Human Behaviour 2021*, 1–7. https://doi.org/10.1038/s41562-021-01098-5

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, *5*(4), 296–320.

Rudolph, A. S. (1994). *Contrast and assimilation effects: A meta-analytic review* (Doctoral dissertation).

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, *76*(5), 805.

Russo, J. E. (2018). Bayesian revision vs. information distortion. *Frontiers in Psychology*, *9*, 1550.

Russo, J. E., Carlson, K. A., Meloy, M. G., & Yong, K. (2008). The goal of consistency as a cause of information distortion. *Journal of Experimental Psychology: General*, *137*(3), 456.

Russo, J. E., Meloy, M. G., & Medvec, V. H. (1998). Predecisional distortion of product information. *Journal of Marketing Research*, *35*(4), 438-452.

Schindler, R. M., & Bickart, B. (2012). Perceived helpfulness of online consumer reviews: The role of message content and style. *Journal of Consumer Behaviour*, *11*(3), 234–243. https://doi.org/10.1002/cb.1372

Schlosser, A. E. (2011). Can including pros and cons increase the helpfulness and persuasiveness of online reviews? The interactive effects of ratings and arguments. *Journal of Consumer Psychology*, *21*(3), 226–239. https://doi.org/10.1016/j.jcps.2011.04.002

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.

Shannon, C. E. (1948). A note on the concept of entropy. *Bell System Tech. J*, *27*(3), 379-423.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as Analysis and Risk as Feelings: Some Thoughts about Affect, Reason, Risk, and Rationality. In *Risk Analysis* (Vol. 24, Issue 2, pp. 311–322). John Wiley and Sons, Ltd. https://doi.org/10.1111/j.0272-4332.2004.00433.x

Statista. (2021). *Online reviews - Statistics and Facts* . https://www.statista.com/topics/4381/online-reviews/

Taylor, J. W. (1974). The Role of Risk in Consumer Behavior. *Journal of Marketing*, *38*(2), 54–60. https://doi.org/10.1177/002224297403800211

Thomas, D. L. (1991). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology*, *59*(2), 291. https://doi.org/10.1037/0022-3514.59.2.291

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

Voss, K. E., Spangenberg, E. R., & Grohmann, B. (2003). Measuring the hedonic and pragmatic dimensions of consumer attitude. *Journal of marketing research*, *40*(3), 310-320.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly journal of experimental psychology*, *20*(3), 273-281.

Watson, D. (1988). Intraindividual and interindividual analysis of positive and negative affect: Their relation to health complaints, perceived stress, and daily activities. *Journal of Personality and Social Psychology*, *54*(6), 1020–1030.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. Psychological bulletin, 98(2), 219.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wundt, W. (1924). *An introduction to psychology* (R. Pintner, Trans.). London: Allen and Unwin. (Original work published 1912)

Yang, C., Lin, K. H. Y., & Chen, H. H. (2009, September). Writer meets reader: Emotion analysis of social media from both the writer's and reader's perspectives. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 287-290). IEEE.

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., & Zhang, H. (2014). Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews. *MIS Quarterly*, *38*(2), 539-560.

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.

Young, J. (2020). US ecommerce sales grow 14.9% in 2019. https://www.digitalcommerce360.com/article/us-ecommerce-sales/

Zhang, J. Q., Craciun, G., & Shin, D. (2010). When does electronic word-of-mouth matter? A study of consumer product reviews. *Journal of Business Research*, *63*(12), 1336-1341.

Zwick, R., Rapoport, A., Lo, A. K. C., & Muthukrishnan, A. V. (2003). Consumer sequential search: Not enough or too much? *Marketing Science*, *22*(4), 503-519.