Seeing Stars: How the Binary Bias Distorts the Interpretation of Customer Ratings

MATTHEW FISHER
GEORGE E. NEWMAN
RAVI DHAR

Across many different contexts, individuals consult customer ratings to inform their purchase decisions. The present studies document a novel phenomenon, dubbed "the binary bias," which plays an important role in how individuals evaluate customer reviews. Our main proposal is that people tend to make a categorical distinction between positive ratings (e.g., 4s and 5s) and negative ratings (e.g., 1s and 2s). However, within those bins, people do not sufficiently distinguish between more extreme values (5s and 1s) and less extreme values (4s and 2s). As a result, people's subjective representations of distributions are heavily impacted by the extent to which those distributions are imbalanced (having more 4s and 5s vs. more 1s and 2s). Ten studies demonstrate that this effect has important consequences for people's product evaluations and purchase decisions. Additionally, we show this effect is not driven by the salience of particular bars, unrealistic distributions, certain statistical properties of a distribution, or diminishing subjective utility. Furthermore, we demonstrate this phenomenon's relevance to other domains besides product reviews, and discuss the implications for existing research on how people integrate conflicting evidence.

Keywords: online user ratings, information integration, binary thinking

Imagine you are on vacation looking for a tasty, local restaurant. Naturally, you might consult several customer-rating websites. Website A shows many restaurants in the area and, beside each one, presents the average customer rating (ranging from 1 to 5 stars). Website B summarizes the same reviewer data, but also reports the frequency of each reviewer score (i.e., the number of 5-star reviews, 4-star reviews, etc.). The present studies seek to answer a simple yet central question: Will people choose a different restaurant after consulting Website A than after consulting Website B?

The results from 10 experiments suggest that, in fact, exposure to a full distribution of scores may actually change people's representations and, as a consequence, their choices. Specifically, we find that people tend to make a categorical distinction between the positive ratings (4s and 5s) and negative ratings (1s and 2s). However, within those bins, people do not sufficiently distinguish between more extreme values (5s and 1s) and less extreme values (4s and 2s). Because of this, people's subjective summary representation of the distribution is impacted by the extent to which the distribution is imbalanced—either top-heavy (more 4s and 5s) or bottom-heavy (more 1s and 2s)—and tends to ignore the extremity of those values. We dub this effect the "binary bias" and demonstrate that it affects people's evaluations and purchase decisions.

The binary bias documented here makes a novel connection between two complementary streams of research.

Matthew Fisher (mcfisher@cmu.edu) is a postdoctoral fellow in social and decision sciences, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. George E. Newman (george.newman@yale.edu) is an associate professor of management and marketing, Yale School of Management, 165 Whitney Avenue, New Haven, CT 06511. Ravi Dhar (ravi.dhar@yale.edu) is the George Rogers Clark Professor of Management and Marketing, Yale School of Management, 165 Whitney Avenue, New Haven, CT 06511. Please address correspondence to Matthew Fisher. The authors acknowledge the helpful input of the editor, associate editor, and reviewers. This article is based on the lead author's dissertation. All raw data files and analytic syntax are available at Open Science Framework (https://osf.io/t3j59/).

Editor: Gita Johar

Associate Editor: Stijn van Osselaer

Advance Access publication Month 0, 0000

 • Vol. 0 • March 2018
DOI: 10.1093/jcr/ucy017

Specifically, we draw upon past research demonstrating people's tendency to rely on simplified heuristics when aggregating information (Gigerenzer, Todd, and the ABC Research Group 1999), as well as work demonstrating consumers' predisposition toward categorical or dichotomous thinking when reasoning about continuous stimuli (Brough and Chernev 2012). Consistent with the notion that the binary bias is driven by dichotomous thinking, we demonstrate that the bias is attenuated when dichotomous cues are removed (i.e., when the scale's labels are changed) and is accentuated when participants are primed with binary as opposed to continuous choices. We also demonstrate that the phenomenon itself is not restricted to certain graphical displays, modes of presentation, or even purchase decisions. For example, we show that the same pattern of results obtains when individuals aggregate transcript grades and make judgments about a student's academic performance.

The remainder of this article is organized as follows: We first review prior work on information integration and binary thinking, which provides the empirical support for predicting the binary bias. Then, across several studies, we use customer ratings as a case study of how the binary bias affects decision making.

THEORETICAL BACKGROUND

Previous research has documented integration inaccuracies across a wide variety of tasks; however, the question of how categorical thinking distorts the summary representation of conflicting evidence has not been directly examined. People do not normatively integrate information across relevant categories to make predictions, but instead tend to consider only the single most likely category (Murphy and Ross 1994). When making intuitive judgments, people inaccurately integrate the extremity of a piece of evidence (proportion of heads in a series of coin flips) with the strength of the evidence (number of total flips) (Griffin and Tversky 1992). And, in some cases, perceptual systems nonnormatively neglect alternative interpretations of ambiguous stimuli (Fleming, Maloney, and Daw 2013). These findings illustrate that the mind often fails to optimally integrate all of the complexities of relevant information and instead utilizes alternative, simplified strategies.

Several such strategies have been identified by previous research. According to Anderson (1981), the way people integrate information reflects a sort of "cognitive algebra," whereby summary representations are formed through simple computations, such as the weighted average. For example, providing mildly positive information alongside highly positive information leads to less favorable responses, suggesting that the evidence is averaged and not added to form a summary judgment (Anderson and Alexander 1971; Troutman and Shanteau 1976). However, other research has found that summary evaluations can be formed implicitly by adding through the addition of weighted evidence (Betsch et al. 2001). In certain cases, individuals seem to not integrate information at all, instead relying on only the single most diagnostic variable, a strategy called the "take-the-best" heuristic (Gigerenzer, Todd, and the ABC Research Group 1999). Computer simulations demonstrate that this simple heuristic can match or even outperform much more complicated models in speed and accuracy (Gigerenzer and Goldstein 1996; Hogarth and Karelaia 2006).

Another related integration strategy is the tallying heuristic (Gigerenzer 2004), where attributes are weighted equally and added up until a threshold is reached. Consumers in low-effort contexts use this type of decision heuristic, as they favor products with more positive features even if those features are of different levels of importance (Alba and Marmorstein 1987). Similarly, participants choose payoff sets with more positive than negative options even if the expected value based on the magnitude of the payoff was lower (Payne et al. 2008).

The Binary Bias

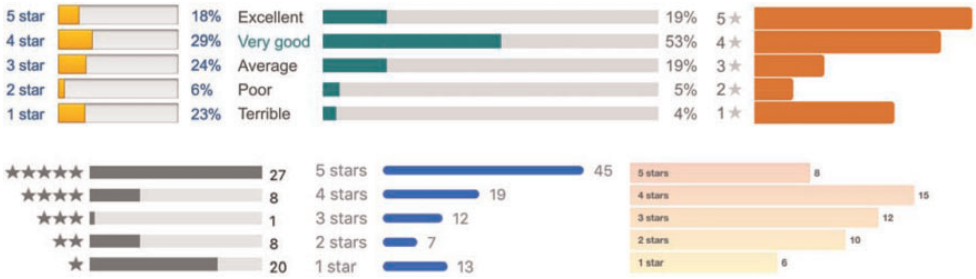
Here we propose one particular process for integrating information, especially information that spans seemingly distinct conceptual categories. We dub this strategy the "binary bias" and define it as the tendency to bin continuous data into discrete categories, such as positive versus negative ratings. Within each of these bins, the total amount of evidence is tallied and people's summary representations reflect the degree to which one category outweighs the other.

Although this process has not been examined in the context of consumer evaluations of products, there is some indirect support for the influence of binary thinking on consumer behavior. In particular, people tend to simplify complex information into discrete categories (Gutman 1982), which then influences their judgments and decisions (Mogilner, Rudnick, and Iyengar 2008). For example, when evaluating food options, people rely on dichotomous categories like healthy/unhealthy and do not sufficiently take into account quantitative aspects (i.e., calories). Thus small amounts of high-calorie food are judged to have more calories than a large amount of low-calorie food (Rozin, Ashmore, and Markwith 1996). Relatedly, an unhealthy food option plus a healthy food option are judged as having fewer calories than the unhealthy food option alone. Foods from opposing categories are averaged together, but when no categorical distinctions are present they are added (Brough and Chernev 2012; Chernev and Gal 2010). This illustrates how treating continuous data as dichotomous can lead to distinct, and sometimes distorted, patterns of reasoning.

We suggest that analogous binary thinking extends to product ratings, which may be intuitively categorized as

FIGURE 1

SAMPLE REVIEW DISTRIBUTIONS FROM AMAZON, TRIPADVISOR, GOOGLE, APPLE, FACEBOOK, AND YELP



positive or negative. For example, when considering a distribution of product reviews, people intuitively distinguish positive ratings (scores above the midpoint) from negative ratings (scores below the midpoint). As a result, consumers may not sufficiently distinguish more extreme values (5s and 1s) from less extreme values (4s and 2s)—these quantities are binned according to their initial categorization. Their tendency to focus on whether the distribution has more positive or negative ratings leads people to insufficiently consider the extremity of those values when forming a summary representation of the data. Thus, the binary bias contributes to an existing literature, which has shown that the initial categorization of stimuli affects its perception. The present studies focus on dichotomous distinctions (positive vs. negative) and specifically how those dichotomous categories drive information integration.

The process of summarizing distributions of scores in a binary manner can be formalized into a single statistic we call the "imbalance score." A distribution's imbalance score equals the difference between the total number of positive ratings (e.g., 5- and 4-star reviews) and the total number of negative ratings (e.g., 1- and 2-star reviews). Thus, the imbalance score reflects the degree to which a given distribution is top-heavy (more positive ratings than negative ratings) or "bottom-heavy" (more negative ratings than positive ratings). This can be contrasted with other possible methods of summarizing data—for example, overweighting the most frequent score (i.e., the mode) or accurately averaging values (i.e., the mean). While the imbalance score can be thought of as another type of summary statistic, it is a psychological construct, as opposed to statistics such as the mean, which is a mathematical construct. The imbalance score provides a way of tracking perceptions based on binary thinking, but we make no claims about its utility as a mathematical concept.

An Illustration: Five-Bar Histograms

Customer reviews provide an ideal test case for the binary bias. Online reviews are an important input for consumer decision making (Chevalier and Mayzlin 2006). Nearly all customer review services (e.g., Amazon.com, TripAdvisor.com, Yelp.com) ask reviewers to provide a numeric score that reflects their assessment of the product or experience. With multiple reviews, this naturally results in a distribution of scores. Rating services often summarize these scores in terms of a measure of central tendency, such as a mean. In many cases, however, consumers can also see the full underlying distribution of scores (see figure 1).

For example, product ratings on Amazon.com are initially presented as a single mean score (ranging from 1 to 5 stars) to allow comparison across products. However, by clicking on a particular product, consumers are presented the full distribution of scores (the percentage of 5-star reviews, 4-star reviews, etc.) as well as the comments provided by each individual reviewer. The distribution of ratings is most often presented graphically as a five-bar histogram. Despite their prevalence and importance, relatively little is known about how people integrate the information provided by these sorts of graphical displays to create a summary representation. As a result, there have been several recent calls within the marketing literature for further research on how online reviews are interpreted and utilized by consumers (Simonson 2015).

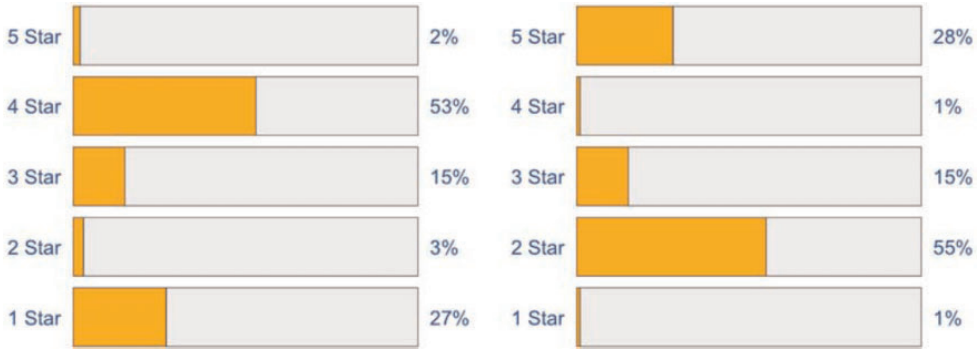
Given the prevalence of five-bar histograms to communicate distributions of customer reviews, the present studies examine how consumers aggregate and interpret those scores. Therefore, we examine five-bar histograms as a case study of potential biases affecting consumer decision making and the integration of information. We also demonstrate, however, that the error in interpreting five-bar histograms reflects a far more general phenomenon with implications that extend beyond graphical displays and purchase decisions.

THE CURRENT STUDIES

In the following studies, we document how the binary bias affects consumer decision making across multiple

FIGURE 2

TOP-HEAVY (LEFT) AND BOTTOM-HEAVY (RIGHT) DISTRIBUTION FROM STUDY 1 ($M = 3$ STARS FOR BOTH DISTRIBUTIONS)



contexts. Specifically, study 1 demonstrates that data sets with identical means may be evaluated very differently based on their underlying distributions. Study 2 replicates this effect and controls for other factors, such as the salience of particular bars in the distribution (i.e., the mode). Study 3 demonstrates the binary bias using an incentive-compatible design and study 4 does so using actual online reviews. Study 5 shows that, in some cases, the binary bias can even lead certain top-heavy distributions with lower true means to be preferred over bottom-heavy distributions with higher true means. Study 6 shows that when the true mean is presented next to the distribution, the bias is still evident, suggesting that exposure to the mean is not sufficient to override the binary bias. Studies 7 and 8 provide a test of the mechanism of categorical thinking and demonstrate that imposing different categorical distinctions on the rating values changes people's summary representations. Study 9 offers converging evidence for the proposed mechanism by showing that after people make dichotomous as opposed to continuous judgments, the strength of the binary bias increases. Finally, study 10 demonstrates that the binary bias generalizes beyond graphical presentations of information and influences other summary estimates, such as estimates of student achievement based on transcript grades.

STUDY 1: EFFECTS ON PRODUCT VALUATION

Study 1 demonstrates that products with identical mean ratings can be valued quite differently depending on the extent to which the underlying distributions are "imbalanced." We presented participants with a series of customer ratings in the form of five-bar histograms modeled after the format used by Amazon.com. The histograms all had a mean rating of 3 stars but differed in the extent to which they were top-heavy (i.e., greater numbers of 4- and 5-star ratings than 1s and 2s) or bottom-heavy (i.e., greater numbers of 1- and 2-star ratings than 4s and 5s) (see figure 2). Participants were asked to rate the products on several measures of valuation (e.g., willingness to pay, purchase intent). We predicted that despite no difference in the true mean ratings, categorical thinking as captured by the imbalance score would lead participants to value products with top-heavy distributions more than products with bottom-heavy distributions.

Method

Participants. Two hundred forty participants (145 male; $M_{AGE} = 33.92$, $SD = 10.87$) from the United States completed the study through Amazon Mechanical Turk (MTurk). Each experiment contained a unique sample of participants, who had not participated in any related studies. Informed consent was obtained from all participants across all experiments.

Materials and Procedure. The stimuli consisted of 40 five-bar histograms, which were randomly selected from all possible distributions, totaling 100% and with a mean of 3 stars ($N = 25, 753$). Each participant viewed a (randomly selected) subset of 10 of those figures. The figures were presented one at a time (in a random order) and, for each one, participants were told that the figure depicted customer ratings for a given product. To enhance the generalizability of the findings, between-subjects we varied the type of product to span a range of small to large purchases. Specifically, participants were told that the ratings were for boxes of candy (small purchase), sets of knives (medium purchase), or cars (large purchase). For each product, participants responded to the following randomly ordered items on a 1–10 Likert scale: "How much would you be willing to pay for this [product]? (Not a Lot–Very Much)," "How likely would you be to buy this [product]? (Very Unlikely–Very Likely)," "How would you expect

your experience of this [product] to be? (Very Negative–Very Positive),” and “How do you feel about this [product]? (Very Unfavorable–Very Favorable).” These four dependent measures were strongly correlated and formed a reliable scale ($\alpha = .94$).

Results and Discussion

Despite identical mean ratings across all of the products, participants' valuation varied dramatically. For the boxes of candy, the scores ranged from 2.28 to 4.87 (SD = .68); for the knives, they ranged from 2.26 to 5.06 (SD = .67); and for the cars, they ranged from 2.09 to 5.97 (SD = .76).

To assess how the distribution itself impacted valuation, we coded each figure in terms of the extent to which it was top-/bottom-heavy. Specifically, we subtracted the total number of 1- and 2-star ratings from the total number of 4- and 5-star ratings. Thus, positive scores reflected top-heavy distributions, while negative scores reflected bottom-heavy distributions. We then conducted a linear mixed-effects regression analysis using the lme4 and lmerTest packages in R (Bates et al. 2015; Kuznetsova, Brockhoff, and Christensen 2015; R Development Core Team 2013), using the imbalance score as a fixed effect. A comparison of models' BIC revealed the product and the product × imbalance interaction term as poor predictors, suggesting that ratings and the effect of the imbalance score were consistent across all product categories, so these were dropped from the model. For random effects, we included intercepts for subjects and items and by-subject random slopes for the effect of imbalance. Since there is no standard for calculating *p*-values in mixed-effects models (Bates et al. 2014), we also computed bootstrapped 95% confidence intervals (CIs) for the coefficients and tested whether these CIs included zero. Throughout the studies, we report standardized and unstandardized coefficients (and standard errors). We found that the extent to which the distribution was imbalanced significantly predicted product valuation ($\beta = .23$, SE = .05, $p < .001$; $b = .04$, SE = .01, 95% CI = [.02, .05]; see figure 3). See table 1 for the full model.

FIGURE 3

BY-ITEM PRODUCT VALUATION AS A FUNCTION OF IMBALANCE SCORE

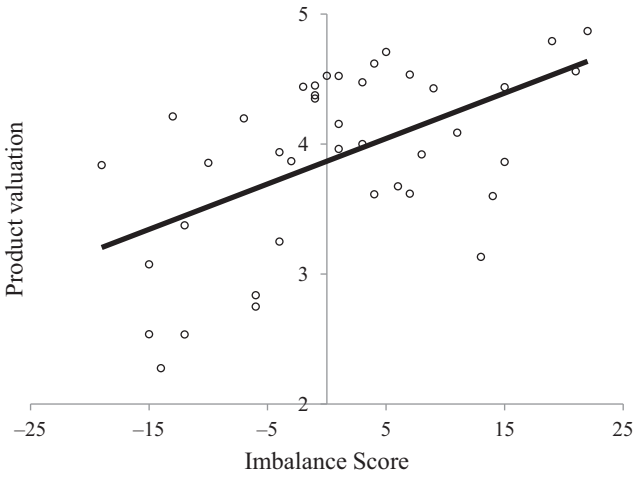


TABLE 1

MIXED-EFFECTS REGRESSION RESULTS FOR VALUATION RATINGS IN STUDY 1

Fixed effects	Estimate *b* (*β*)	SE	Bootstrapped 95% CI	
(Intercept)	3.87 (.00)	.13 (.07)	3.64	4.08
Imbalance	**.04 (.23)**	**.01 (.05)**	**.02**	**.05**

Random effects	Predictor variable		SD
Grouping variable: Subject		Intercept	1.10
	Imbalance	Slope	.03
Grouping variable: Item		Intercept	.52

NOTE.— Observations: 2,400; subjects: 240; items: 40.

STUDY 2: CONTROLLING FOR GRAPHICAL FEATURES

Study 1 indicated that, despite having the same mean, products with top-heavy ratings were valued by consumers more than products with bottom-heavy ratings. It is possible that even though the means were held constant, the difference could arise due to other features of the figures. Thus, the aim of study 2 was to address potential alternative explanations as well as explore downstream consequences of binary thinking.

It could be that other statistical features of the distributions are driving the effect. For example, participants could simply focus on the most frequent rating (i.e., the mode). If their attention is drawn to the highest bar, then those ratings could be disproportionately weighted and shift their valuations. Additionally, given that the median of an array of numbers can influence judgment (Parducci et al. 1960; Smith, Diener, and Wedell 1989), a distribution's median could be predictive of participants' valuations. We also tested if the standard deviation or kurtosis (“peakedness”) of the distributions explained participants' responses. To differentiate between these accounts, we presented participants with the same stimuli as in study 1. We also asked them to report the bar that most captured their attention. Our account predicts that imbalance scores will have an effect on valuation when we control for the effect of the most salient bar as well as other statistical features.

Finally, we tested if the binary bias affects estimates of the means as well as subjective evaluations. This comparison is interesting because there is obviously an actual mean value for each distribution. Therefore, if these estimates are biased we are able to quantify the extent to which they differ from the true mean.

Method

Participants. Eighty participants (55 male; $M_{AGE} = 36.20$, SD = 10.60) from the United States completed the study through MTurk.

Materials and Procedure. Study 2 followed the same procedure as study 1, except that: a) since there were no item differences found in study 1, participants considered ratings only for one product category (cars); and b) participants were asked two additional questions: "Based on your immediate judgment, on average, how many stars did this product receive?" on a 1–5 sliding scale, which could be adjusted to the hundredths decimal place. Participants were asked to use their "immediate judgment" to discourage them from actually calculating the true mean. Second, participants reported, "Which bar in the graph most captures your attention?" on a 1–5 Likert scale.

Results and Discussion

As in study 1, the four dependent variables formed a highly reliable scale ($\alpha = .97$). We conducted a linear mixed-effects regression with product valuation as the outcome variable and the imbalance score as the predictor. The most salient bar (self-reported attention), statistical mode, standard deviation, nonparametric skew (Arnold and Groeneveld 1995), and kurtosis were included as covariates. Since the median and parametric skew were significantly correlated with imbalance score, they were not included in the analysis. Each participant viewed a different subset of distributions, so each subject's average imbalance score across the 10 items was also included as a control variable. Intercepts for subjects and items, and slopes for the by-subject effect of imbalance, were included as random effects. Imbalance scores, $\beta = .20$, SE = .04, $p < .001$; $b = .03$, SE = .01, 95% CI = [.02, .05], and the most salient bar, $\beta = .32$, SE = .03, $p < .001$; $b = .46$, SE = .04, 95% CI = [.37, .52], significantly predicted product valuation. See table 2 for the full results of the regression analysis.

Additionally, participants' estimates of the mean ($M = 2.99$, SD = .54) were predicted by imbalance score, $\beta = .27$, SE = .04, $p < .001$; $b = .01$, SE = .00, 95% CI = [.01, .02]. This result suggests that the binary bias not only extends to summary representations that impact consumer valuations, but also influences downstream statistical judgments. This result shows the effect is a bias in that participants' estimates of the mean deviated from the true value.

TABLE 2

MIXED-EFFECTS REGRESSION RESULTS FOR VALUATION RATINGS IN STUDY 2

Fixed effects	Estimate b (β)	SE	Bootstrapped 95% CI	
(Intercept)	2.97 (.00)	.70 (.08)	1.63	4.49
Imbalance	**.03 (.18)**	**.01 (.03)**	**.02**	**.04**
Attention	**.46 (.32)**	**.04 (.03)**	**.37**	**.54**
Mode	.07 (.05)	.13 (.09)	−.18	.34
Standard deviation	−.18 (−.03)	.27 (.04)	−.81	.34
Nonparametric skew	.10 (.05)	.19 (.09)	−.24	.42
Kurtosis	−.04 (−.02)	.09 (.04)	−.23	.13
Subset imbalance	.03 (.04)	.05 (.08)	−.06	.13

Random effects	Predictor variable	SD
Grouping variable: Subject	Intercept	.65
Imbalance	Slope	.06
Grouping variable: Item	Intercept	.06

NOTE.— Observations: 790; subjects: 79; items: 40.

Lastly, we analyzed whether the undersensitivity to the difference between the 1- and 2-bar differed from the undersensitivity to the difference between the 4- and 5-bar. We conducted the same linear mixed-effects model as above, but replaced the predictor of imbalance with fixed effects for the low bars (1-bar + 2-bar) and high bars (4-bar + 5-bar). As expected, the effect of low-end bars is negative, $\beta = -.30$, SE = .06, $p < .001$; $b = -.04$, SE = .01, 95% CI = [−.05, −.02], and the effect of high-end bars is positive, $\beta = .27$, SE = .06, $p < .001$; $b = .02$, SE = .01, 95% CI = [.01, .04], but furthermore, we found little difference in their predictive strength, indicating that participants are influenced by both the high end and low end of the ratings, $t(154) = 1.04$, $p = .30$.

Importantly, the results of study 2 rule out the alternative account that participants' self-reported attention to a particularly salient bar solely explains the variance in responses. It shows that other statistical features of the distributions are not driving the effect. However, this study could not rule out an effect of the median or parametric skew—a point addressed in studies 7–9. Additionally, study 2 extends the previous finding to statistical judgments, demonstrating that the effect is a bias. And lastly, the results suggest the bias is symmetric in that the negative and positive reviews are weighted roughly equally.

STUDY 3: CONSEQUENTIAL PURCHASES

In studies 1 and 2, participants rated products in hypothetical purchase scenarios. To increase external validity, the aim of study 3 was to demonstrate the binary bias in an incentive-compatible context. Will people still show the

binary bias when real money is at stake as they make their decisions?

Method

Participants. One hundred twenty participants (53 male; $M_{AGE} = 33.72$, SD = 10.43) from the United States completed the study through MTurk.

Materials and Procedure. Participants reported their willingness to pay (WTP) for a random subset of 10 of 40 world music albums. They were asked, "How many cents are you willing to pay for this album that has received the following reviews" (0–500 cents) and then viewed a distribution of 1- to 5-star ratings for that album. The set of 40 distributions used in study 3 was the same set used in studies 1 and 2.

To make the study incentive-compatible, we adopted a double-lottery BDM procedure (Becker, DeGroot, and Marschak 1964; Fuchs, Schreier, and van Osselaer 2015). At the beginning of the study, participants were told that after they made 10 WTP ratings, the experimenter would pool all of the purchase decisions from all participants and randomly select some of them to actually happen. If one of their purchases was selected, their WTP would be compared against a randomly selected price. If their WTP was greater than or equal to the random price, they would pay that amount for a download link for that album and would also receive the remainder of their $5. If their maximum WTP was less than the random price, they would not receive the album and would receive the $5. At the end of the study, decisions were selected and participants were paid and or received the download link in the manner specified in the instructions.

Results and Discussion

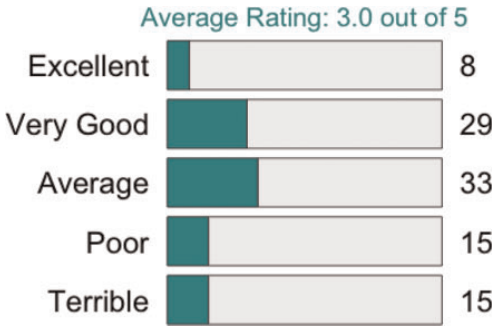
We conducted a linear mixed-effects regression with imbalance score as a predictor of participants' WTP, including random intercepts for subjects and items and random slopes for the by-subject effect of imbalance. WTP ratings were square-root transformed to address right skewness and zero values in the data. Replicating the results of the previous studies in an incentive-compatible context, we found that imbalance scores were a significant predictor of participants' willingness to pay, $\beta = .06$, SE = .03, $p = .03$; $b = .03$, SE = .02, 95% CI = [.01, .06]. This result shows that the binary bias affects consumer decision making when real money is at stake.

STUDY 4: REAL-WORLD RATINGS

Studies 1–3 demonstrated the binary bias using artificially constructed stimuli designed so that the imbalance score varied while the true mean remained constant. Study 4 aimed to replicate the effect using distributions taken from an actual online rating website. By using real-world ratings, we did not confine the distributions to the statistical properties of our artificial selection process; instead, properties like correlations between certain ratings and variance in imbalance reflected their natural occurrence. Specifically, participants rated their willingness to stay at hotels after considering those hotels' ratings from the travel review website TripAdvisor.com.

FIGURE 4

SAMPLE ITEM FROM STUDY 4



Method

Participants. One hundred twenty participants (63 male; $M_{AGE} = 37.08$, SD = 11.70) from the United States completed the study through MTurk.

Materials and Procedure. We compiled distributions of ratings for all hotels in the city of Los Angeles with an average customer rating of 3 out of 5 stars from TripAdvisor.com ($N = 43$). To match how people would be presented ratings when actually evaluating hotels, the average rating was displayed above the ratings distribution. Additionally, the color scheme and labels matched those from the original source (see figure 4).

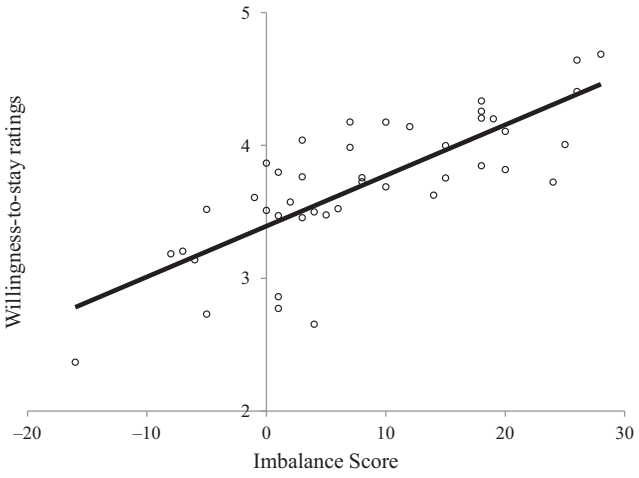
Participants were instructed that they would view a variety of hotel ratings for a town they would be visiting soon. For each distribution, participants were asked, "How willing would you be to stay at this hotel?" on a sliding scale from 1 (Not at all) to 7 (Very) with hundredth decimal place precision. Using a random sampling method, we had each participant view a random subset of 15 (of 43) distributions.

Results and Discussion

Replicating the results of the previous studies, we found evidence for the binary bias when participants considered actual hotel ratings. A linear mixed-effect model with random intercepts for items and subject plus random slopes for the by-subject effect of imbalance on ratings found that

FIGURE 5

BY-ITEM WILLINGNESS-TO-STAY RATINGS AS A FUNCTION OF IMBALANCE SCORE



imbalance scores significantly predicted participants' ratings, $\beta = .29$, SE = .04, $p < .001$; $b = .04$, SE = .005 (see figure 5). This suggests not only that the binary bias contributes to the theoretical understanding of how people integrate statistical information, but also that this process has an impact on how people consider real-world consumer ratings.

STUDY 5: EVALUATIONS OF MEANS VERSUS DISTRIBUTIONS

In studies 1–4, participants viewed customer ratings in isolation. In many real-world contexts, however, people compare product ratings, which can lead to different modes of processing (Hsee et al. 1999). Therefore, in study 5, we examined if the binary bias influences how people choose between multiple offerings.

In particular, we were interested in contexts where evaluations based on means might importantly differ from evaluations based on the distributions. Following the example in the introduction, participants were presented with two forms of rating summaries for restaurants: means and five-bar histograms. In the means condition, participants viewed only the mean ratings—one restaurant had a slightly higher mean than the other (e.g., 3.15 vs. 3.00). In the five-bar histograms condition, as in studies 1–4, participants were presented with the distributions underlying those means (the true means were not displayed). The pairs of distributions were constructed such that the restaurant with a lower mean had a top-heavy distribution, while the restaurant with a higher mean had a bottom-heavy distribution. We expected that when provided with only means, people should (unsurprisingly) choose the restaurant with the higher average rating. However, when presented with the distributions, we tested whether the binary bias would lead participants to instead prefer the restaurant with the top-heavy distribution (with lower true mean) over a bottom-heavy distribution (with a higher true mean).

Method

Participants. Two hundred participants (129 male; $M_{AGE} = 33.19$, SD = 9.65) from the United States completed the study through MTurk.

Materials and Procedure. Participants viewed four pairs of reviews and for each pair were asked, "Which restaurant would you prefer?" The reviews for one restaurant in each pair had a mean of 3.00 and the other restaurant's reviews had a mean of 3.10, 3.15, 3.20, or 3.25. In the means condition, participants were presented with only the average rating. In the distributions condition, participants were presented with only the distributions underlying those means. The pairs of distributions were constructed such that the lower-rated restaurant's (3.00) distribution was top-heavy and the higher rated restaurant's (3.10, 3.15, 3.20, or 3.25) distribution was bottom-heavy. In the means condition, each participant viewed each of the four possible pairs of averages. In the distributions condition, each participant viewed one of two possible pairs for each mean value, making a total of four choices. See the appendix for details of the distributions used in this study.

Results and Discussion

We ran a logistic regression, with information (average vs. distribution) and combinations (3.00 vs. 3.10; 3.00 vs. 3.15; 3.00 vs. 3.20; 3.00 vs. 3.25) as predictors of participants' preference for the restaurant with a higher mean. Participants chose the higher mean option less often when they viewed the distributions, $b = -3.82$, SE = .29, $p < .001$. In pairs with more similar true means, participants were more likely to choose the item with the lower mean, $b = 4.91$, SE = 1.75, $p = .005$. See table 3 for a summary of the results. These results demonstrate that viewing averages versus distributions can lead products with lower mean ratings to be preferred over products with higher mean ratings.

STUDY 6: EVALUATION IN PRESENCE OF MEANS AND DISTRIBUTIONS

Studies 1–5 demonstrated that distributions of ratings can shift preferences when no additional statistical information is provided. However, when viewing distributions of ratings in the real world, consumers are often provided with the mean alongside the distribution. In study 6, we tested if participants' subjective evaluations would still be

TABLE 3

MEAN PREFERENCE FOR LOWER-MEAN OPTION IN STUDY 5

	3.25 versus 3.00	3.20 versus 3.00	3.15 versus 3.00	3.10 versus 3.00
Mean only	3%	2%	1%	8%
Distribution only	58%	51%	64%	71%

influenced by the binary bias, even when the mean was readily available.

Method

Participants. Three hundred twenty-two participants (206 male; $M_{AGE} = 33.36$, SD = 10.57) from the United States completed this study through MTurk.

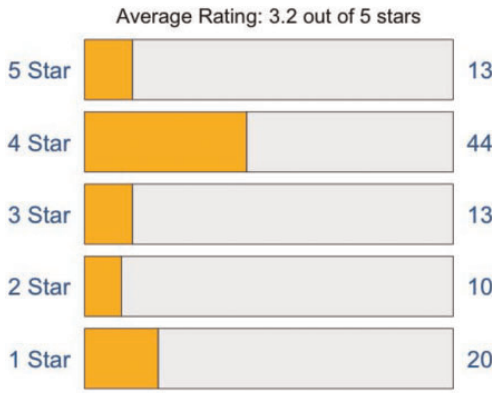
Materials and Procedure. In study 6, participants were asked to imagine that they were visiting a town in the near future. They were told that they would be viewing a distribution of ratings for 15 restaurants in that town. Each participant viewed 15 top-heavy or 15 bottom-heavy ratings distributions. Each set of 15 consisted of restaurants with means of 3.2, 3.5, 3.8, 4.1, and 4.4. To create the stimuli set used in this study, we generated a random selection of 40 distributions for each of the five mean values. The three most top-heavy and three most bottom-heavy distributions of each set of 40 were used. Critically, the true mean of both sets was identical. A midpoint of 3.8 was selected since it is the mean restaurant rating on the popular restaurant review website Yelp.com.

As participants viewed the ratings for each restaurant, the true mean of each distribution was clear, with the following label placed above each graph: "Average Rating: [X] out of 5 stars" (see figure 6). The distributions were presented to participants one at a time in a randomized order. For each distribution, participants rated how willing they would be to try the restaurant on a Likert scale from 1 (Not at all) to 7 (Very much). After viewing all 15 restaurant's ratings, participants were asked, "How excited would you be to eat at the restaurants in this town?" (1 [Not at all] to 7 [Very]), "How likely would you be to try the restaurants in this town?" (1 [Not at all] to 7 [Very]), and "What is your impression of the quality of the restaurants in this town" (1 [Very low] to 7 [Very high]). These three measures were combined to form a composite measure of liking. Finally, participants were asked to "please estimate the average review rating for all the restaurants you just viewed" on a 1–5 sliding scale with their response shown to the hundredths decimal place.

To enhance the salience of the mean, we asked half of the participants to actually write the mean rating for each trial. Participants in these conditions could advance to the next page only if their answer matched the reported mean.

FIGURE 6

SAMPLE STIMULI FROM STUDY 6



Results and Discussion

A linear mixed-effects model predicted willingness-to-try ratings using imbalance and display condition (write mean vs. do not write mean), including random intercepts for subject and item and by-item random slopes for imbalance, display, and imbalance × display interaction. Replicating the results of study 5, the model revealed imbalance as a significant predictor. See table 4 for the complete model. Using a likelihood ratio test, we compared this model's goodness of fit to a second identical model, which also included the imbalance × display interaction term as a fixed effect. This test revealed no significant difference between the models, $\chi^2 = .25$, $df = 1$, $p = .62$, suggesting that writing out the mean did not affect participants' willingness-to-try ratings.

We next assessed participants' liking judgments for the set of restaurants they viewed. A linear regression, using imbalance and display to predict liking ratings, again found a significant effect of imbalance (low), $\beta = -.44$, SE = .11, $p < .001$; $b = -.42$, SE = .11. Using a likelihood ratio test, we compared this model's goodness of fit to the same model that also included the imbalance × display interaction term. This test revealed no significant difference between the models, $F(1, 318) = .90$, $p = .34$, again suggesting that writing out the mean for each distribution did not change their valuation of the set of restaurants.

Lastly, we analyzed participants' mean estimates as assessed by the mean memory judgments at the end of the study. A linear regression predicted mean estimates using imbalance and display, and found a significant effect of imbalance, $\beta = -.37$, SE = .11, $p < .001$; $b = -.14$, SE = .04. We then compared the goodness of fit to a model including the imbalance × display interaction term, using a likelihood ratio test. This test revealed a significant

TABLE 4

MIXED-EFFECTS REGRESSION RESULTS FOR WILLINGNESS-TO-TRY RATINGS IN STUDY 6

Fixed effects	Estimate b (β)	SE	Bootstrapped 95% CI	
(Intercept)	4.94 (.03)	.21 (.15)	4.52	5.40
Imbalance (low)	**−.27 (−.20)**	**.10 (.08)**	**−.49**	**−.04**
Display (do not write mean)	.11 (.08)	.09 (.07)	−.07	.29

Random effects	Predictor variable		SD
Grouping variable: Subject		Intercept	.74
Grouping variable: Item		Intercept	.82
	Imbalance	Slope	.16
	Display	Slope	.07
	Imbalance × display	Slope	.25

NOTE.— Observations: 4,830; subjects: 322; items: 15.

imbalance × display interaction, $F(1, 318) = 4.43$, $p = .04$, such that when participants did not write out the mean, their memory for the top-heavy distributions ($M = 3.83$, SD = .38) was higher than for the bottom-heavy distributions ($M = 3.59$, SD = .46), but when they wrote out the mean, their memory for the high-imbalance restaurants ($M = 3.73$, SD = .34) was no different than their memory for the low-imbalance restaurants ($M = 3.68$, $SD = .34$). Together, the results from study 6 show that unsurprisingly, when the mean is especially salient, mean estimates more accurately reflect the true mean. Nonetheless, the salience of the mean did not change participants' ratings, suggesting that summary representation of the distributions independently influences people's subjective evaluations. In other words, when the full distribution is presented, it does not appear that the mean is sufficient to override the binary bias. Furthermore, these results suggest that in consumer contexts where the true mean is displayed alongside review distributions, we would expect the binary bias to persist.

STUDY 7: BIVALENT VERSUS UNIVALENT RATINGS

The aim of study 7 was to provide a direct test of the psychological mechanism of dichotomous thinking. The central claim of the binary bias is that the difference in subjective evaluations arises because people dichotomize a continuous scale into positive and negative scores. This account predicts that if the scale was not perceived as dichotomous, but rather as a continuous dimension, then the preferences resulting from imbalanced distributions should be attenuated.

To test this, we utilized the same distributions from study 5. However, we manipulated each bar's corresponding label. In one condition, the bars were labeled to suggest a dichotomous range of values (Very Poor–Very Good), while in the other condition, they were labeled to suggest a univalent range of values (Fair–Extremely Good). See figure 7 for sample stimuli.

By using categorization cues instead of the shape of the distribution to elicit the binary bias, study 7 clarifies the process underlying the effect. One alternative account addressed by this study is that the skewness of the distributions are driving participants' ratings (Mitton and Vorkink 2007). In study 2, our analysis controlled for the mean (first moment), standard deviation (second moment), and kurtosis (fourth moment), but did not include parametric skewness (third moment), because of its strong correlation with the imbalance score. Median was excluded from the analysis for the same reason. If the skewness or the median, not binary thinking, is driving the pattern of results in the earlier studies, then there would be no difference based on whether the labels of the distribution are dichotomous or univalent. If, however, categorical thinking underlies the binary bias, then we would expect participants' preference for top-heavy distribution to be weaker when dichotomous cues are removed.

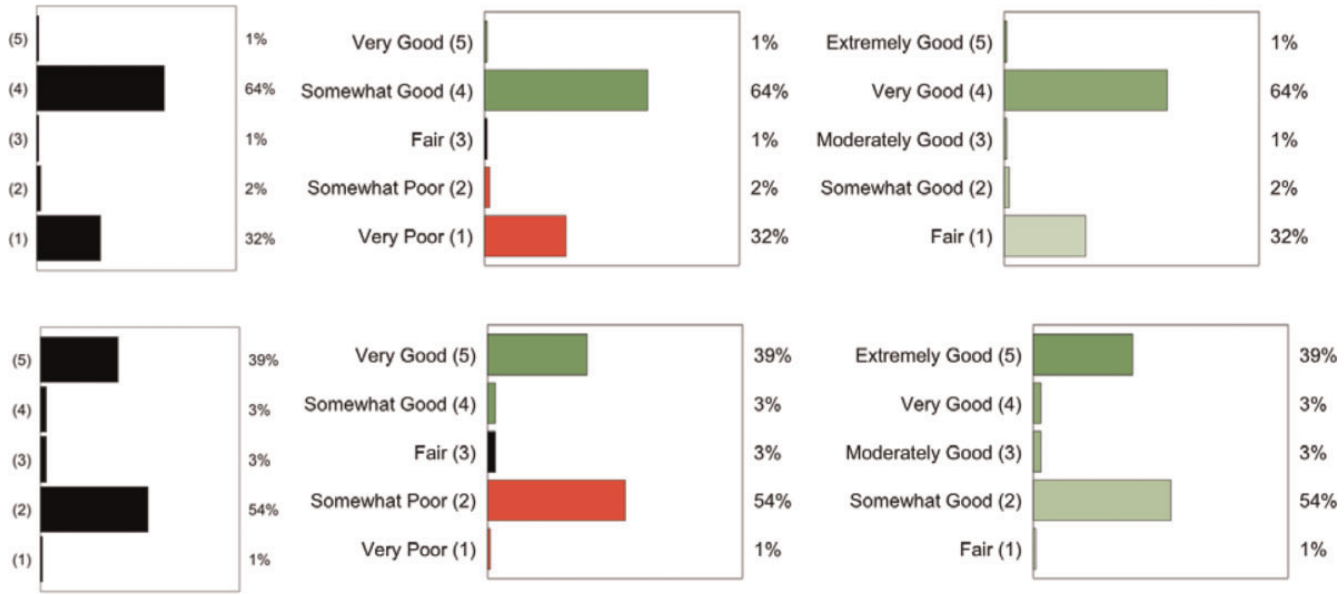
Method

Participants. Two hundred participants (123 male; $M_{AGE} = 32.41$, SD = 9.46) from the United States completed the study through MTurk. The baseline condition was conducted separately with one hundred one participants (45 male; $M_{AGE} = 33.88$, SD = 10.72).

Materials and Procedure. Participants were randomly assigned to the bivalent, univalent, or control condition. Participants across all conditions viewed both versions of one of the four combinations of distributions used in study 5 (e.g., pair 1a and pair 1b from the appendix). Again, participants were asked, "Which car would you prefer to purchase?" In the bivalent condition, the y-axis categories were labeled from 1 (Very Poor) to 5 (Very Good), the 1- and 2-bars were colored red, the 3-bar was colored black, and the 4- and 5-bars were colored green. In the univalent

FIGURE 7

SAMPLE STIMULI FROM THE CONTROL (COLUMN 1), BIVALENT (COLUMN 2), AND UNIVALENT (COLUMN 3) CONDITIONS IN STUDY 7



condition, the y-axis categories were labeled from 1 (Fair) to 5 (Extremely Good) and all five bars were colored green with lower-value bars colored lighter shades. In the control condition, no verbal labels were provided and all bars were colored black (see figure 7).

Results and Discussion

To test the effect of binary presentation, we ran a mixed-effects logistic regression, with labels (bivalent vs. univalent vs. control) and combinations (3.00 vs. 3.10; 3.00 vs. 3.15; 3.00 vs. 3.20; 3.00 vs. 3.25) as fixed effects, random intercepts for subjects and items, and random slopes for the by-item effect of labels on ratings. As predicted by our categorical thinking account, participants' preference for the lower-rated option with a higher imbalance score was weaker when the reviews were presented with univalent labels as opposed to bivalent labels, $b = -1.39$, $SE = .43$, $p = .002$, and control, $b = -.94$, $SE = .39$, $p = .02$. There was no significant difference between the control and bivalent, $b = .44$, $SE = .39$, $p = .26$. There was also a significant main effect of combinations, $b = -5.63$, $SE = 2.81$, $p = .04$, as participants were more willing to select the lower-rated car when the difference between the true means was smaller. See table 5 for a summary of the results. Consistent with our theory, the control condition patterned nearly identically to the bivalent condition, suggesting that participants naturally interpret the histograms in terms of binary categories. Furthermore, these results demonstrate that removing a salient conceptual midpoint can attenuate the binary bias, suggesting that the underlying effect is explained by a tendency to bin evidence into conceptually discrete categories, not by any particular statistical feature of the data (e.g., skewness, median).

TABLE 5

MEAN PREFERENCE FOR LOWER-MEAN OPTION IN STUDY 7

	3.25 versus 3.00	3.20 versus 3.00	3.15 versus 3.00	3.10 versus 3.00
Control	63%	74%	72%	62%
Bivalent	63%	79%	69%	82%
Univalent	46%	48%	50%	71%

STUDY 8: CATEGORICAL THINKING

The results of studies 1–7 provide robust evidence for the proposed binary bias. However, a plausible alternative account of how people integrate individual ratings could also explain the evidence presented thus far. If people's subjective value of the ratings scale is S-shaped, then the pattern of results from the previous studies could arise because of an underweighting of extreme points (1- and 5-star ratings) relative to less extreme points (2- and 4-star ratings). In study 7, it is possible that participants engaged

in subjective discounting only when a salient midpoint was present. Since ratings were shifted into the positive domain in the univalent condition, categorization cues were confounded with midpoint presence. Participants could have engaged in subjective discounting in the bivalent condition but not in the univalent condition since there was no midpoint. Thus, study 7 was unable to rule out diminishing marginal value as a possible explanation.

To test binary thinking against subjective weighting, we manipulated the degree to which a given distribution was categorized into dichotomous bins, without removing the midpoint. In study 8, the shape of six-point distributions was held identical across conditions, but the histogram was grouped into one category (baseline), two categories, or three categories. This allowed us to test if the influence of particular bars changed based on how they are categorized. For example, do participants rate a distribution more favorably when the tall 4-bar is included in the high category than when it is included in the medium category? If so, it would suggest that people's interpretation of the data is driven by categorical thinking as opposed to a diminishing subjective weighting of the positive and negative side of the scale.[1]

Method

Participants. Four hundred eighty participants (262 male; $M_{AGE} = 35.89$, SD = 11.46) from the United States completed the study through MTurk.

Materials and Procedure. Participants viewed six-bar rating distributions for a random subset of 15 (of 40) restaurants. Participants were assigned to the one-category (baseline), two-category, or three-category condition. In the baseline condition, all six bars were colored black and the y-axis was labeled with numbers only. In the two-category condition, the top three bars were colored green and labeled "High," and the bottom three bars were colored red and labeled "Low." In the three-category condition, the top two bars were colored green and labeled "High," the middle two bars were colored black and labeled "Medium," and the bottom two bars were colored red and labeled "Low." See figure 8 for sample stimuli from each condition. Since the baseline condition did not include any verbal labels, all participants were told at the beginning of the study that the restaurants had been rated on a scale from 1 (Lowest) to 6 (Highest).

We created the 40 distributions used in study 8 by randomly generating distributions with the 3- or 4-bar as the tallest bar. Creating distributions with this property led to large differences between imbalance scores in the two-category condition (sum of 5s, 6s, and 7s minus sum of 1s, 2s, and 3s) and imbalance scores in the three-category condition (sum of 5s and 6s minus sum of 1s and 2s). Some distributions had a mode of 3 so that negative reviews were influential in the two-category condition, and others had a mode of 4 so that positive reviews were influential in the two-category condition. Thus, distributions with a mode above and below the midpoint were evenly represented in the stimuli. In the two-category condition, the restaurants with a mode of 4 had an average imbalance score of 14.4, and the restaurants with a mode of 3 had an average imbalance score of −7.9. When the exact same distributions are split into three categories, the restaurants with a mode of 4 had an average imbalance score of −21.7, and the restaurants with a mode of 3 had an average imbalance score of 25.6. For example, if participants attend to binary distinctions, this shift can be seen in the stimuli in figure 8 (row 2): the imbalance score in the two-category condition equals –2 and in the three-category condition equals –23. Thus, across conditions the visual cues change the category to which certain bars belong and shift the imbalance of that distribution. Note, there were baseline differences in the true mean of the mode = 4 ($M = 3.26$) and mode = 3 ($M = 3.85$) restaurants, so mode = 3 restaurants were expected to be rated higher in the baseline (one-category) condition.

For each restaurant, participants were asked, "How willing would you be to try this restaurant?" and responded on a sliding scale from 1 (Not at all) to 7 (Very). The binary bias predicts that the categorization cues should change willingness-to-try ratings—flipping the preference for mode = 3 and mode = 4 items between the two-category and three-category conditions. However, alternative accounts that rely on a particular statistical feature (e.g., skewness) or differential weighting across different bars would predict no difference across the three conditions.

Results

A linear mixed-effects regression analyzed the relationship between categorization and willingness-to-try ratings. Category (baseline vs. two-category vs. three-category) and mode (4 vs. 3) were included as fixed effects, without the interaction term. Random intercepts for subjects and items, and random slopes for the by-item effect of category and the by-subject effect of mode, were also included. Using a likelihood ratio test, we compared this model's goodness of fit to a separate model that was identical but also included the category × mode interaction term. This comparison suggested a significant interaction, $\chi^2 = 103.90$, $df = 2$, $p < .001$. The results of the second model showed that compared to the baseline condition, the preference for restaurants with a mode of 3 over restaurants with a mode of 4 became stronger in the three-category condition. But in the two-category condition (with less extreme imbalance scores), the preference flips: restaurants with a mode of 4 are rated higher than restaurants with a mode of 3 (see figure 9 and table 6 for results of the regression analysis). These results demonstrate that even when the heights of the bars are consistent across distributions,

1 We thank an anonymous reviewer for suggesting this experiment.

FIGURE 8

SAMPLE STIMULI FROM THE BASELINE (COLUMN 1), TWO-CATEGORY (COLUMN 2), AND THREE-CATEGORY (COLUMN 3) CONDITIONS IN STUDY 8

categorization cues can alter how that information is interpreted.

Discussion

In study 8, participants show a "trinary" bias by creating a neutral bin (3–4) in addition to a positive (5–6) and negative bin (1–2). This is supported by the baseline condition responding more similarly to the three-category condition than the two-category condition. In fact, in the studies using five-bar distributions, participants show a similar pattern by differentiating the neutral bar (3) from the positive and negative bar. However, we conceptualize the effect as a creation of two categories around a midpoint, thus the *binary* bias. We favor this terminology because the main predictor we propose, the imbalance score, does not take into account the midpoint bar(s). Thus, the neutral bin is essentially ignored as people compare dichotomous categories.

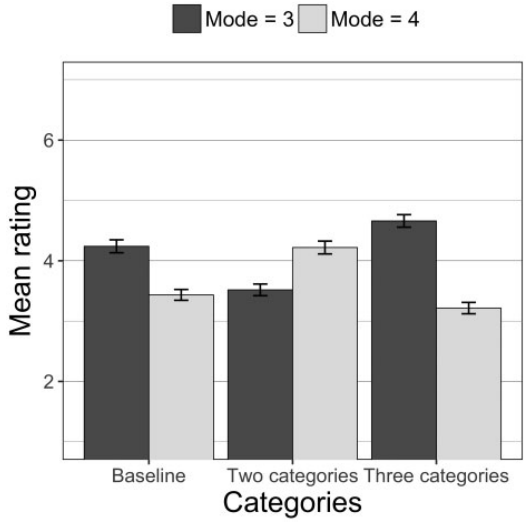
Together, studies 7 and 8 provide a critical test of the process underlying the binary bias. In these studies, categorization cues led to changes in the influence of certain data points even though the actual distributions remained identical across conditions. In the previous studies, we operationalized binary thinking by carefully constructing stimuli to have identical true means but different imbalance scores. Even though we included many other control variables in our analyses, this study design left open the possibility that some other statistical feature of the distributions could be explaining the results. For example, skewness and median, which are highly correlated with imbalance, or an S-shaped subjective weighting could be the actual mechanism. Studies 7 and 8 provide strong evidence against these alternative accounts. When binary thinking is induced through categorization cues, we find differences in valuation that cannot be explained by any particular statistical feature since the distributions under consideration are otherwise identical.

STUDY 9: PRIMING CATEGORICAL THINKING

Studies 1–6 demonstrated the binary bias by manipulating the imbalance distributions, while studies 7 and 8 did so by altering the presentation format. Study 9 tested the proposed mechanism in a third way: priming categorical thinking by asking participants to first make dichotomous as opposed to continuous judgments. If people are more reliant on the imbalance score after having made categorical judgments, it would be strong evidence that categorical thinking helps explain how people are summarizing online ratings.

FIGURE 9

WILLINGNESS-TO-TRY RATINGS BY CATEGORY AND MODE IN STUDY 8 (ERROR BARS, MEAN ± STANDARD ERROR)



Method

Participants. Three hundred fifty-two participants (185 male; $M_{AGE} = 35.37$, SD = 10.80) from the United States completed the study through MTurk.

Materials and Procedure. Study 9 took place in two phases: the prime phase and the test phase. Participants were randomly assigned to either the binary or continuous condition. In the prime phase, all participants were presented with a random subset of 10 (of 40) of the car rating distributions from study 1 and asked, "If you were considering buying a car with these customer ratings, how would you rate this car?" In the binary condition, participants replied with a forced choice (Good or Bad), while those in the continuous condition replied on a sliding scale from 0 (Bad) to 100 (Good) that showed participants their response to the hundredths decimal place. In the test phase, participants were then asked the same four consumer valuation questions from study 1 for an additional random subset of 10 (of 40) of the same car rating distributions from the prime phase.

Results and Discussion

In line with previous studies, a linear mixed-effects regression model with random intercepts for subjects and items, and slopes for the by-subject effect of imbalance on ratings, showed that the imbalance score predicted participants' valuations. Using a likelihood ratio test, we compared this model's goodness of fit to a second identical model, which also included the imbalance × condition interaction term as a fixed effect. This test revealed a significant difference between the models, $\chi^2 = 5.00$, $df = 1$, $p = .03$, indicating that those in the binary condition were

TABLE 6

MIXED-EFFECTS REGRESSION RESULTS FOR WILLINGNESS-TO-TRY RATINGS IN STUDY 8

Fixed effects	Estimate b (β)	SE	Bootstrapped 95% CI	
(Intercept)	4.23 (.27)	.15 (.12)	3.96	4.50
Category (two category)	**−.75 (.55)**	**.15 (.11)**	**−1.03**	**−.43**
Category (three category)	**.42 (.31)**	**.15 (.11)**	**.12**	**.70**
Mode (mode = 4)	**−.80 (−.58)**	**.18 (.13)**	**−1.16**	**−.45**
Two category × positive	**1.51 (1.11)**	**.15 (.11)**	**1.18**	**1.79**
Three category × positive	**−.65 (−.47)**	**.14 (.11)**	**−.89**	**−.36**
Random effects		Predictor variable		SD
Grouping variable: Subject			Intercept	.85
	Mode = 4		Slope	.76
Grouping variable: Item			Intercept	.49
	Two category		Slope	.21
	Three category		Slope	.12

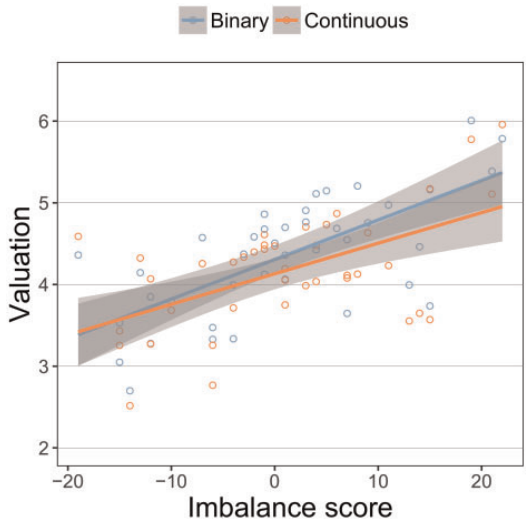
NOTE.— Observations: 3,600; subjects: 240; items: 40.

more reliant on the imbalance score than those in the continuous condition (see figure 10 and table 7).

Though it was not a planned analysis, there was an effect of the manipulation on top-heavy distributions but not bottom-heavy distribution ($p = .002$). The Johnson-Neyman technique showed that the effect of condition on valuation was significant for imbalance scores above −3.65. While this result could very well be a statistical fluke, it raised the question as to whether the relationship between imbalance and valuation was driven only by top-heavy distributions. To address this issue, we conducted a meta-analysis of all studies where valuation judgments were elicited for both types of distributions (studies 1–3). We find that there is a strong effect of imbalance for top-heavy distributions ($\beta = .09$, $SE = .02$, $p < .001$) and for bottom-heavy distributions ($\beta = .14$, $SE = .02$, $p < .001$). This indicates that while the effectiveness of the prime in study 9 may interact with the valence of the imbalance score, dichotomization plays a role in summarizing distributions regardless of valence.

FIGURE 10

THE RELATIONSHIP BETWEEN VALUATION AND IMBALANCE BY CONDITION IN STUDY 9



STUDY 10: EXTENDING THE BINARY BIAS TO OTHER DOMAINS

We next explored the generality of the phenomenon. The previous studies all used similar graphical displays to examine the binary bias. However, the effect itself is hypothesized to be a dichotomization of information more generally—an interpretation that is strongly supported by the results of studies 7 and 8. The aim of study 10 was to test the binary bias in a new domain using a completely different presentation of data. Accordingly, participants viewed transcripts and rated students' academic performance. The transcripts presented letter grades as raw data (see figure 11). As with the distributions from the previous studies, we calculated imbalance scores by splitting the data at the midpoint, subtracting the total Ds and Fs from the total As and Bs for each transcript. We hypothesized that a distribution's imbalance score would predict participants' GPA estimates and ratings of academic achievement.

Method

Participants. Two hundred participants (101 male; $M_{AGE} = 35.07$, $SD = 11.04$) from the United States completed the study through MTurk.

Materials and Procedure. Twenty-four transcripts were used as stimuli in study 10. The 24 transcripts were randomly selected from all possible combinations of 15 grades that averaged to a C, with the constraint that at least one set of grades for each possible imbalance score (−5 to +5) was selected. Each participant viewed a random subset of 15 of the 24 transcripts. They were asked, "Please estimate the GPA of this student" on a Likert scale from 0 (F) to 4 (A) and "How would you assess the academic achievement of this student?" on a sliding scale from 0 (Very Poor)

TABLE 7

MIXED-EFFECTS REGRESSION RESULTS FOR VALUATIONS IN STUDY 9

Fixed effects	Estimate b (β)	SE	Bootstrapped 95% CI	
(Intercept)	4.30 (.05)	.11 (.07)	4.09	4.55
Imbalance	**.05 (.28)**	**.01 (.05)**	**.03**	**.06**
Condition (continuous)	−.17 (−.11)	.12 (.07)	−.39	.05
Imbalance × condition	**−.01 (−.07)**	**.00 (.03)**	**−.02**	**.00**

Random effects	Predictor variable		SD
Grouping variable: Subject	Imbalance	Intercept	1.06
		Slope	.03
Grouping variable: Item		Intercept	.49

NOTE.— Observations: 3,520; subjects: 352; items: 40.

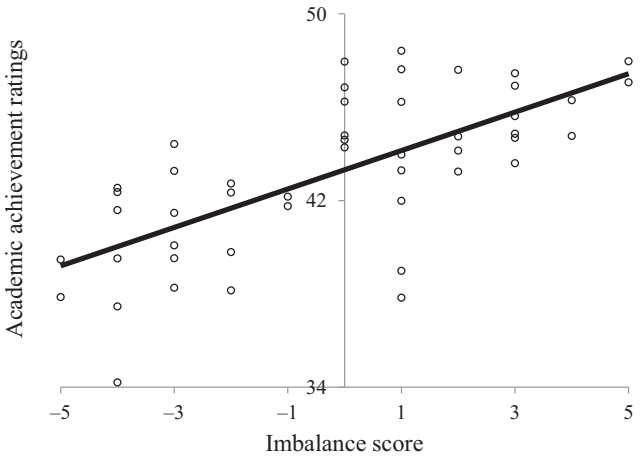
FIGURE 11

SAMPLE STIMULI FROM STUDY 9 WITH AN IMBALANCE SCORE = −2, MODE = D, AND TRUE MEAN = C

Course Name	Grade
Europe and Empire	A
Social Psychology	A
Intro to Statistics	A
Vector Calculus	A
Microeconomics	A
History of Rome	B
Consumer Behavior	C
The Cold War	D
Intro to Programming	D
Organic Chemistry	D
Constitutional Law	D
Basics of Astrophysics	D
Intro to Typography	F
Molecular Biology	F
Advanced Spanish	F

FIGURE 12

THE RELATIONSHIP BETWEEN IMBALANCE AND ACADEMIC ACHIEVEMENT RATINGS IN STUDY 10



to 100 (Very Good). Additionally, participants were randomly assigned to receive the grades in a descending order (As to Fs) or in random order. This factor, however, did not affect the results and therefore we collapsed across this dimension when examining the effect of imbalance on GPA estimates and ratings of academic achievement.

Results and Discussion

We conducted a linear mixed-effects regression with academic achievement ratings as the outcome variable and imbalance score and mode as the predictors. Random intercepts for subjects and items, and slopes for the by-subject effect of imbalance on ratings, were also included in the model. Imbalance score was a strong predictor of academic achievement ratings, $\beta = .16$, SE = .05, $p = .001$; $b = .03$, SE = .01, 95% CI = [.01, .04] (see figure 12). In line with the previous studies, mode was also a significant predictor, $\beta = .10$, SE = .04, $p = .02$; $b = .04$, SE = .02, 95% CI = [.03, .05]. Additionally, we tested if imbalance score also affected GPA estimates using the same fixed and random effects as the previous model. Again we found an effect of imbalance, $\beta = .13$, SE = .04, $p = .001$; $b = .61$, SE = .17, 95% CI = [.29, .99], and an effect of mode, $\beta = .09$, SE = .03, $p = .02$; $b = .99$, SE = .41, 95% CI = [.25, 1.80]. Similar to study 2, this result suggests that the shape of the distribution affects not only consumer-related judgments, but more abstract statistical estimates as well.

As when judging products based on consumer reviews, participants in study 10 displayed a binary bias when evaluating students based on their transcripts. Importantly, this replication suggests that the results from the previous studies are not due to idiosyncratic features of how people conceptualize five-star rating scales. This suggests that the binary bias is a domain-general heuristic, affecting how data is summarized across a variety of contexts.

GENERAL DISCUSSION

The present studies document a novel phenomenon, the binary bias, and its effects on consumer decision making. In short, we find that when viewing summary distributions of product reviews (such as five-bar histograms), people consider the relative number of positive versus negative ratings, and underweight the extremity of the scores within each category. This process of integrating evidence alters the perceived mean of the distribution as well as people's purchasing decisions. This effect can give rise to paradoxical cases in which products with lower mean ratings are preferred over products with higher mean ratings. In addition, showing the true mean of the distribution does not counteract the bias, as evidenced by participants' judgments from memory. Critically, these findings are driven by categorical thinking, as demonstrated by the shift in consumers' valuation when different grouping labels are used to describe identical distributions of reviews. We further demonstrated our proposed mechanism by priming people to think categorically, and finding that they show a greater reliance on the imbalance score. Finally, we document that the binary bias occurs outside of a consumer decision-making context, indicating that it may reflect a domain-general process.

More generally, our claim is not that imbalance is the only way in which consumers form summary representations of data. In fact, we identified other relevant factors in the current studies, such as the mode and particularly

salient bars. Rather, we aim to document one way—that is, the binary bias—that is conceptually interesting for reasons having to do with categorical thinking. We demonstrate the role of categorical thinking in a manner that is not as readily explained by alternatives such as skewness, the median, or S-shaped weighting functions (e.g., studies 7–9). That said, the manner in which people form subjective impressions based on ratings is certainly multiply determined.

Theoretical Implications

These studies contribute to the understanding of an important psychological question: How does the mind summarize conflicting evidence? Specifically, the binary bias highlights the way in which categorical logic pervades the mind. From high-level social-cognitive processes (Macrae and Bodenhausen 2000; Park and Rothbart 1982) to low-level visual processes (Fleming et al. 2013), quantitative information is often compressed into a qualitative format. The current studies suggest that the summary of evidence occurs in a similar manner; information-rich evidence is simplified into a binary representation.

Furthermore, the analyses used in these studies—operationalizing the binary bias as an imbalance score—could be used to measure the degree to which binary thinking occurs across a wide variety of domains. Notably, in study 10, not only did participants reason about a context different from the previous studies, but the information was presented in a very different, nongraphical format. Nonetheless, we found strong relationships between the imbalance score and participants' responses. This suggests that the binary bias is not confined to the ways in which people interpret graphs, but could offer a more general theory of information integration.

This raises the question as to why people integrate information in this way. One possible reason is that discounting the extremity of evidence makes the task of integrating a range of values cognitively tractable. We have limited cognitive resources, and simplifying the computational complexity may be the more efficient solution. Thus, the binary bias could be an example of the mind satisficing instead of optimizing (Simon 1982). The binary bias reduces complexity more than other proposed heuristics. For example, some research has suggested that people compute a weighted average of the available evidence (Anderson 1981), a process that requires a weight and a value for each piece of evidence. The binary bias, however, assigns only one of two values (such as positive or negative) and weighs each piece of evidence equally. Given that people bin data to simplify the process of integration, they are quite accurate at utilizing this heuristic, as shown by the strong relationship between imbalance scores and valuation.

Marketing Implications and Future Directions

Customer ratings and reviews are a key component of the current consumer environment. Reviews have been shown to impact perceptions of quality (Aaker and Jacobson 1994) and predict sales across a variety of product categories (Chevalier and Mayzlin 2006; Ye, Law, and Gu 2009). Although customer reviews are not always genuine (Mayzlin, Dover, and Chevalier 2014) and do not align with independent rating agencies, such as *Consumer Reports* (De Langhe, Fernbach, and Lichtenstein 2016), 70% of consumers report trusting online consumer reviews, second only to recommendations from family and friends (92%; Nielsen 2012). Despite the importance of customer reviews and their corresponding ratings, relatively little work has investigated how people naturally interpret them.

The current studies show that displaying a five-bar histogram as opposed to the mean can lead to very different outcomes. We find that strategies designed to give customers helpful additional information can alter their choices. Participants in our studies are not basing their estimates of the mean by mathematically averaging the data provided; instead, they are using a cognitive shortcut that leads to systematically biased estimates. This suggests that marketers should be cautious in using graphical depictions to summarize important information. Even clear labels like those used in study 6 are not enough to counteract the effects of binary processing. In other words, graphical depictions that may seem intuitive can be easily misinterpreted.

While our analyses focused on the influence of the binary bias, it is also worth noting that the salient bars independently influenced consumers' valuation. This is another example of how low-level features of a graphical display can distort how data is interpreted (Fischer 2000; Graham 1937; Stone et al. 2003). Recent studies converge on the idea that the shape of the distribution of reviews influences consumer decision making. People are more willing to tolerate dispersive reviews when the diversity of tastes in the product domain is greater (He and Bond 2015). Further, bimodal rating distributions are preferred when a product expresses a personal taste (Rozenkrants, Wheeler and Shiv 2017). There may be cases where these phenomena and the binary bias are both relevant; for example, high self-expression could lead consumers to prefer a bimodal distribution, while the binary bias might make the same bimodal distribution be viewed less favorably. Based on studies 7–9, which shifted valuation without altering any features of the distributions themselves, we do not see these findings as potential explanations of the binary bias. Instead, as previously mentioned, the effect of graphical displays on consumer preferences is certainly multiply determined. In the current set of studies, for example, we show that salience and imbalance score both independently influence consumers' valuations. Furthermore, given that we find the binary bias within domains of personal preference (e.g., musical albums) as well as outside (e.g., set of knives), it is likely that these motivational accounts are orthogonal to our primarily cognitive account: binary thinking. Future work could explore cases

where multiple cues, such as imbalance, diversity of taste, and multimodality, are pitted against each other.

In clinical psychology, dichotomous thinking has been linked to perfectionism (Egan et al. 2007) and increased emotional reactions to the self and others (Epstein and Meier 1989). Perfectionists tend to engage in dichotomous thinking, which leads to underperformance in consumer decision-making tasks because they prematurely abandon problems with no single, ideal answer (He 2016). These individual differences may also influence the degree to which people exhibit the binary bias. Future work could examine whether individuals with a tendency to think dichotomously (Byrne et al. 2008) more readily ignore the relative weight of evidence when integrating information.

Conclusion

Understanding the factors that influence the evaluation of customer ratings is an important yet relatively understudied topic in marketing. The binary bias helps explain how such reviews are integrated and summarized. These findings offer insights into multiple consumer contexts, particularly online review platforms, but appear to apply to information integration more generally. At the very least, the current studies may prevent you from choosing a potentially inferior restaurant on your next vacation.

DATA COLLECTION INFORMATION

The first author collected and analyzed all data from spring 2016 to spring 2018 using Amazon Mechanical Turk under the supervision of the second and third authors.

APPENDIX

Distribution-Only Condition Stimuli in Study 5

	1 star	2 stars	3 stars	4 stars	5 stars	True mean	Imbalance score
Pair 1a	1	54	3	3	39	3.25	–13
	32	2	1	64	1	3.00	31
Pair 1b	1	56	1	1	41	3.25	–15
	30	4	4	60	2	3.00	28
Pair 2a	1	57	1	3	38	3.20	–17
	31	3	2	63	1	3.00	30
Pair 2b	1	57	2	1	39	3.20	–18
	30	5	2	61	2	3.00	28
Pair 3a	1	58	2	3	36	3.15	–20
	31	4	1	62	2	3.00	29
Pair 3b	1	59	1	2	37	3.15	–21
	29	7	1	61	2	3.00	27
Pair 4a	1	60	2	2	35	3.10	–24
	32	2	2	62	2	3.00	30
Pair 4b	1	61	1	1	36	3.10	–25
	29	6	2	62	1	3.00	28

REFERENCES

Aaker, David A. and Robert Jacobson (1994), "The Financial Information Content of Perceived Quality," *Journal of Marketing Research*, 31 (2), 191–201.

Alba, Joseph W. and Howard Marmorstein (1987), "The Effects of Frequency Knowledge on Consumer Decision Making," *Journal of Consumer Research*, 14 (1), 14–25.

Anderson, Norman H. (1981), *Foundations of Information Integration Theory*, San Diego, CA: Academic Press.

Anderson, Norman H. and Gwendolyn R. Alexander (1971), "Choice Test of the Averaging Hypothesis for Information Integration," *Cognitive Psychology*, 2 (3), 313–24.

Arnold, Barry C. and Richard A. Groeneveld (1995), "Measuring Skewness with Respect to the Mode," *American Statistician*, 49 (1), 34–8.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2014), "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, 67 (1), 1–48.

—— (2015), "lme4: Linear Mixed-Effects Models Using Eigen and S4," R package version 1.1–12, https://CRAN.R-project.org/package=lme4.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak (1964), "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, 9 (3), 226–32.

Betsch, Tilmann, Henning Plessner, Christiane Schwieren, and Robert Gütig (2001), "I Like It but I Don't Know Why: A Value-Account Approach to Implicit Attitude Formation," *Personality and Social Psychology Bulletin*, 27 (2), 242–53.

Brough, Aaron R. and Alexander Chernev (2012), "When Opposites Detract: Categorical Reasoning and Subtractive Valuations of Product Combinations," *Journal of Consumer Research*, 39 (2), 399–414.

Byrne, Susan M., Karina L. Allen, Emma R. Dove, Felicity J. Watt, and Paula R. Nathan (2008), "The Reliability and Validity of the Dichotomous Thinking in Eating Disorders Scale," *Eating Behaviors*, 9 (2), 154–62.

Chernev, Alexander and David Gal (2010), "Categorization Effects in Value Judgments: Averaging Bias in Evaluating Combinations of Vices and Virtues," *Journal of Marketing Research*, 47 (4), 738–47.

Chevalier, Judith A. and Dina Mayzlin (2006), "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43 (3), 345–54.

De Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 42 (6), 817–33.

Egan, Sarah J., Jan P. Piek, Murray J. Dyck, and Clare S. Rees (2007), "The Role of Dichotomous Thinking and Rigidity in Perfectionism," *Behaviour Research and Therapy*, 45 (8), 1813–22.

Epstein, Seymour and Petra Meier (1989), "Constructive Thinking: A Broad Coping Variable with Specific Components," *Journal of Personality and Social Psychology*, 57 (2), 332–50.

Fischer, Martin H. (2000), "Do Irrelevant Depth Cues Affect the Comprehension of Bar Graphs?" *Applied Cognitive Psychology*, 14 (2), 151–62.

Fleming, Stephen M., Laurence T. Maloney, and Nathaniel D. Daw (2013), "The Irrationality of Categorical Perception," *Journal of Neuroscience*, 33 (49), 19060–70.

Fuchs, Christoph, Martin Schreier, and Stijn M. J. van Osselaer (2015), "The Handmade Effect: What's Love Got to Do with It?" *Journal of Marketing*, 79 (2), 98–110.

Gigerenzer, Gerd (2004), "Fast and Frugal Heuristics: The Tools of Bounded Rationality," in *Blackwell Handbook of Judgment and Decision Making*, ed. Derek J. Koehler and Nigel Harvey, Malden: Blackwell, 62–88.

Gigerenzer, Gerd and Daniel G. Goldstein (1996), "Reasoning the Fast and Frugal Way: Models of Bounded Rationality," *Psychological Review*, 103 (4), 650–69.

Gigerenzer, Gerd, Peter M. Todd, and the ABC Research Group (1999), *Simple Heuristics That Make Us Smart*, New York: Oxford University Press.

Graham, James L. (1937), "Illusory Trends in the Observations of Bar Graphs," *Journal of Experimental Psychology*, 20 (6), 597–608.

Griffin, Dale and Amos Tversky (1992), "The Weighing of Evidence and the Determinants of Confidence," *Cognitive Psychology*, 24 (3), 411–35.

Gutman, Jonathan (1982), "A Means-End Chain Model Based on Consumer Categorization Processes," *Journal of Marketing*, 46 (2), 60–72.

He, Stephen X. and Samuel D. Bond (2015), "Why Is the Crowd Divided? Attribution for Dispersion in Online Word of Mouth," *Journal of Consumer Research*, 41 (6), 1509–27.

He, Xin (2016), "When Perfectionism Leads to Imperfect Consumer Choices: The Role of Dichotomous Thinking," *Journal of Consumer Psychology*, 26 (1), 98–104.

Hogarth, Robin M. and Natalia Karelaia (2006), ""Take-the-best" and other simple strategies: Why and when they work "well" with binary cues," Theory and Decision, 61 (3), 205–49.

Hsee, Christopher K., George F. Loewenstein, Sally Blount, and Max H. Bazerman (1999), "Preference Reversals between Joint and Separate Evaluations of Options: A Review and Theoretical Analysis," *Psychological Bulletin*, 125 (5), 576–90.

Kuznetsova, Alexandra, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen (2015), "Package 'lmerTest,'" R Package Version 2.0.

Macrae, C. Neil and Galen V. Bodenhausen (2000), "Social Cognition: Thinking Categorically about Others," *Annual Review of Psychology*, 51 (1), 93–120.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014), "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104 (8), 2421–55.

Mitton, Todd and Keith Vorkink (2007), "Equilibrium Underdiversification and the Preference for Skewness," *Review of Financial Studies*, 20 (4), 1255–88.

Mogilner, Cassie, Tamar Rudnick, and Sheena S. Iyengar (2008), "The Mere Categorization Effect: How the Presence of Categories Increases Choosers' Perceptions of Assortment Variety and Outcome Satisfaction," *Journal of Consumer Research*, 35 (2), 202–15.

Murphy, Gregory L. and Brian H. Ross (1994), "Predictions from Uncertain Categorizations," *Cognitive Psychology*, 27 (2), 148–93.

Nielsen (2012), "Consumer Trust in Online, Social and Mobile Advertising Grows," http://www.nielsen.com/us/en/insights/news/2012/consumer-trust-in-online-social-and-mobile-advertising-grows.html.

Parducci, Allen, Robert C. Calfee, Louise M. Marshall, and Linda P. Davidson (1960), "Context Effects in Judgment: Adaptation Level as a Function of the Mean, Midpoint, and Median of the Stimuli," *Journal of Experimental Psychology*, 60 (2), 65–77.

Park, Bernadette and Myron Rothbart (1982), "Perception of Out-Group Homogeneity and Levels of Social Categorization: Memory for the Subordinate Attributes of In-Group and Out-Group Members," *Journal of Personality and Social Psychology*, 42 (6), 1051–68.

Payne, John W., Adriana Samper, James R. Bettman, and Mary F. Luce (2008), "Boundary Conditions on Unconscious Thought in Complex Decision Making," *Psychological Science*, 19 (11), 1118–23.

R Development Core Team (2013), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Rozenkrants, Bella, S. Christian Wheeler, and Baba Shiv (2017), "Self-Expression Cues in Product Rating Distributions: When People Prefer Polarizing Products," *Journal of Consumer Behavior*, 44 (4), 759–77.

Rozin, Paul, Michele Ashmore, and Maureen Markwith (1996), "Lay American Conceptions of Nutrition: Dose Insensitivity, Categorical Thinking, Contagion, and the Monotonic Mind," *Health Psychology*, 15 (6), 438–47.

Simon, Herbert A. (1982), *Models of Bounded Rationality*, Cambridge, MA: MIT Press.

Simonson, Itamar (2015), "Mission (Largely) Accomplished: What's Next for Consumer BDT-JDM Researchers," *Journal of Marketing Behavior*, 1 (1), 9–35.

Smith, Richard H., Edward Diener, and Douglas H. Wedell (1989), "Intrapersonal and Social Comparison Determinants of Happiness: A Range-Frequency Analysis," *Journal of Personality and Social Psychology*, 56 (3), 317–25.

Stone, Eric R., Winston R. Sieck, Benita E. Bull, J. Frank Yates, Stephanie C. Parks, and Carolyn J. Rush (2003), "Foreground: Background Salience: Explaining the Effects of Graphical Displays on Risk Avoidance," *Organizational Behavior and Human Decision Processes*, 90 (1), 19–36.

Troutman, C. Michael and James Shanteau (1976), "Do Consumers Evaluate Products by Adding or Averaging Attribute Information?" *Journal of Consumer Research*, 3 (2), 101–6.

Ye, Qiang, Rob Law, and Bin Gu (2009), "The Impact of Online User Reviews on Hotel Room Sales," *International Journal of Hospitality Management*, 28 (1), 180–2.