Binary binning: examining the mapping between continuous and binary review scales

WORKING PAPER: 29th Sep 2025

Neel Ocean[1*], Vasundhara[1], Rucha Paricharak[1]

[1] University of Warwick, Gibbet Hill Road, CV4 7AL, UK.

* Corresponding author

E-mail: neel.ocean@warwick.ac.uk

Acknowledgements

Neel Ocean thanks Kajim Hussein for pilot study data collection. Ethical approval for the pilot study was granted by BSREC at the University of Warwick. Ethical approval for the main studies was granted by HSSREC via the Psychology Department at the University of Warwick. Data for all studies is available at: [redacted]

Abstract

Products and services are usually rated on either a five-point scale or a binary scale (i.e. positive vs negative). Using a pilot study and two between-subjects experiments, this paper investigates how individuals evaluate products differently depending on the scale used, how they implicitly categorise ratings on a five-point scale into binary bins, and how they estimate five-point distributions from binary scales. Individuals perceive products as higher in quality when ratings are presented on a binary scale, assuming reviews have been assigned to positive or negative categories based on whether they are above or below the midpoint of a five-point scale. Individuals perceive products as being of equivalent quality across scales only when ratings of four and five are taken as positive, and the remainder as negative. However, when individuals are asked to generate a five-point ratings distribution from binary ratings, they do not account for this skewed perception of positivity *unless* the five-point scale is labelled so that the 'neutral' point of the scale is defined as four rather than three. These findings have theoretical implications for understanding how people implicitly categorise and bin ratings, as well as practical implications for industry and policymakers.

Keywords: consumer behaviour, e-commerce, product ratings, binary-bias, judgement and decision-making

Introduction

Online ratings that communicate information about product quality are in widespread use across many different consumer domains. Products are typically evaluated on one of two main scales: (1) a binary scale, where individuals rate only whether something is 'good' or 'bad'; or (2) a multi-point scale, most commonly a five-point or five-star scale. However, it is not clear how product evaluations translate between the two scales. In particular, where do individuals draw the line between a 'good' rating and a 'bad' rating on a multi-point scale? Furthermore, how do consumers value a product that has the same ratings but summarised on different scales? Understanding this appears to be crucial if we want to understand how consumers make decisions from product rating information. The present research seeks to answer this question by running two main between-subjects experiments. First, we determine how individuals bin ratings from a five-point scale into a binary scale to arrive at product evaluations that are equivalent across scales. Second, we explore the five-point rating distributions individuals generate when shown product ratings on a binary scale to determine whether the implied underlying distribution suggests that people are aware of the tendency to bin ratings in an asymmetric way.

Previous literature has assessed how individuals evaluate products from reviews and ratings, as well as the impact of ratings on purchasing behaviour. In general, people have the tendency to moan or brag about products (Hu et al., 2006), though there may be subtle cultural differences in rating behaviour (Koh et al., 2010). As such, extreme ratings are frequently perceived as more useful, possibly because they convey a less confusing signal about quality (Park & Nicolau, 2015), though the extent to which this is true depends on the type of product as well as distributional characteristics (Mudambi & Schuff, 2010; Lee et al., 2021). Ratings are crucial in determining product choice and subsequent purchasing decisions. For example, Luca (2016) found that a one-star increase in Yelp ratings led to a 5%-9% increase in restaurant revenue. In general, higher valence, higher volume, and lower variance of ratings are associated with more positive consumer responses and a greater likelihood of purchase, though these effects can be moderated by factors like product type and initial valence (Chevalier & Mayzlin, 2006; Chintagunta et al., 2010; Sun, 2012; Langan et al., 2017; Etumnu et al., 2020). Consumers also appear to trust negative reviews more than positive reviews, unless the volume of positive reviews is particularly large (Sparks & Browning, 2011; Gavilan et al., 2018), which is likely to be due to a negativity bias in how ratings are perceived (Baumeister et al., 2001; Rozin & Royzman, 2001; Ocean, 2024). The significance of product ratings has continued to grow due to the increased prevalence of online shopping, especially since the Covid-19 pandemic (Gu et al., 2021).

Online purchasing often involves processing a large amount of information before making a final decision. It has been well established that people use heuristics to simplify the decision making process when cognitive processing becomes too effortful or is otherwise constrained (see Hjeij & Vilks, 2023 for a review). One such heuristic is categorical thinking, which helps to reduce the complexity of decisions by reducing cognitive load and facilitating quicker decision-making (e.g. Gutman, 1982; Mogilner et al., 2008). For example, doctors often simplify complex symptoms and indications into a small number of broad groups (e.g. sick vs normal) so that treatment decisions are easier to make (Elstein & Schwarz, 2002); and investors group firms into distinct 'styles' to make investment decisions despite the varied nature of corporate performance (Barberis & Shleifer, 2003). One specific form of categorisation that is particularly relevant to ratings is *binary bias*. People exhibit a tendency to simplify signals that come from a range of possible values into two groups that indicate two distinct evaluations (Fisher et al., 2018; Fisher & Keil, 2018). When they studied five-point distributions, as are common in product ratings, Fisher et al. (2018) found that people seem to take midpoint ratings (i.e. 3) as neutral, then bin ratings above this as 'good' (i.e. 4 and 5) and ratings below this as 'bad' (i.e. 1 and 2). Therefore, when faced with a multi-point rating scale, it appears that people essentially group ratings into two categories to simplify decision making.

Perhaps in implicit acknowledgement of this phenomenon, some online platforms have implemented binary rating systems that aim to simplify both the review process and the interpretation of reviews. For example, YouTube utilises a thumbs-up/thumbs-down system for video ratings, while Netflix formerly used a similar approach for content recommendations. One comparison of binary and five-star rating systems found that although the five-star scale leads to longer decision-making time, users appear to derive greater satisfaction and intention to use the scale (Chen, 2017), which may explain its relative ubiquity in online retail. However, not all previous research agrees with this. It has been argued that binary ratings are preferable to multi-point scales because they are more likely to eliminate decision-making biases, i.e. the natural System 1 tendency for people to categorise translates more realistically to binary ratings (Harvey, 2016).

The usage of both five-point and binary rating scales in practice leads to a natural question: how do evaluations of the same product with the same ratings differ when they are presented on different scales? Do people use a symmetric binning heuristic to interpret five-point ratings in the spirit of Fisher et al. (2018)? Dehaene's (2011) research on numeric cognition also appears to support a midpoint split. He found that people can envision a five-point scale spatially, and that adults tend to perceive numerical differences relatively uniformly across a scale. Hence, treating the mathematical midpoint of the range as the inflection point between good and bad appears to be a natural dividing

point on a mental number line. However, other work suggests that in the case of online ratings in particular, people do not view positive and negative signals in a uniform or symmetric way. Recent work by Ocean (2024) suggests that an asymmetric weighting may be being applied to ratings, with people valuing any rating of three or higher as an equally positive signal, a rating of one as a negative signal, and a rating of two as somewhere in between. This suggests that, in contrast to Fisher et al. (2018), the 'neutral' score on a five-point scale may be seen as two rather than three. On the other hand, the fact that it takes stronger or a higher volume of 'good' signals to offset 'bad' ones (Baumeister et al., 2001) may suggest that individuals only view very high ratings as positive signals. This implies that the neutral category may even be seen as four. Some empirical evidence supports this idea. Participants who selected "yes" on a binary response scale for a question most frequently selected the fourth point rather than the highest fifth point when asked the same question on a five-point Likert scale, while participants who chose the "no" option responded more neutrally (second and third points) on the five-point scale (Dolnicar & Grün, 2013). This result suggests that individuals may be more likely to interpret undecidedness or neutrality as a negative response in binary terms. The idea that the perceived neutral point is above the scale midpoint is also implicit in a customer recommendation metric known as the Net Promoter Score (Reichheld, 2011), in which scores of seven or eight over a 10-point range are considered neutral.

Other models of judgement also suggest a neutral point that may lie away from the scale midpoint. Parducci's (1965) range-frequency theory finds that judgments are based on both where a signal falls within the scale range, as well as the relative frequency of signals across the range (i.e. how the signals are distributed). This suggests that a person's interpretation of a rating as positive or negative depends on both its numerical value as well as on the observed distribution of ratings. Therefore, the point at which consumers split ratings into 'good' and 'bad' categories may not be fixed, but may instead shift depending on the context and overall distribution of ratings they encounter. While this means that ratings may not have a stable and non-contextual interpretation, individuals may categorise *a priori* based on the ratings distributions they have been exposed to and form a reference point from which to base judgements, in a similar way to decision by sampling (Stewart et al., 2006). Empirical evidence suggests most rating distributions are J-shaped, i.e. bimodal with a large peak at five-stars and a smaller peak at one-star (Hu et al., 2006, 2009). Theoretically, this is due to the tendency for consumers to only post reviews when they are either extremely satisfied or extremely dissatisfied because leaving a review incurs a small cost (Lafky, 2014). If consumers are desensitised to high ratings from repeated exposure, this suggests that the perceived neutral point of a five-point scale lies above the midpoint rather than below it to compensate for the high volume of positive signals. In sum, the literature offers different predictions on where consumers draw the line

between 'good' and 'bad' ratings. Therefore, the first goal of the present study is to understand which ratings are categorised as positive or negative.

However, some websites additionally attach verbal labels to each point of the rating scale. This is likely to prime individuals in how they categorise ratings. For example, TripAdvisor attaches the labels "Amazing", "Very Good", "Average", "Poor" and "Terrible" to ratings of 5, 4, 3, 2, and 1 respectively. Previous research has found that labelling scales affects judgements. When different labels are used for the same point rating, the distributions obtained from each scale are different (Klockars & Yamagishi, 1988). Shifting an 11-point scale down from a midpoint of 5 to a midpoint of 0 reduced the proportion of responses to a subjective life evaluation that were in the lower half of the scale by 21 percentage points (Schwarz et al., 1991). Attaching neutral labelling to a point that is above the midpoint attenuates ratings away from the extremes, while adding emotive labels to the endpoints has the opposite effect (Tsekouras, 2017). Fisher et al. (2018) showed participants two five-point rating distributions for cars and asked them which car they preferred. The distribution were shown with either no labelling, bivalent labelling ('very bad' to 'very good'), or univalent labelling ('fair' to 'extremely good'). They found that preferences under bivalent labelling appear to closely correspond with no labelling, but the absence of a neutral point in univalent labelling causes participants to tend towards the product that has the higher mean rating. When ratings scales are skewed by labelling (either with more negative than positive labels, or with more positive than negative labels), people attach more importance to the label than the numeric scale point value (Wildt & Mazis, 1978). Furthermore, a scale with a predominance of negative labels has more impact than a scale with a predominance of positive labels because people are more reluctant to use the negative categories. Therefore, the ratings will be biased upwards (Wildt & Mazis, 1978). Overall, the findings on scale labelling suggest that binning is dependent on conceptual frameworks and not statistical properties of the distribution, i.e. binary categorisation depends on where the perceived neutral point is (Fisher et al., 2018).

While it is clear that scale labelling affects rating behaviour, we still do not fully understand whether people can translate binary ratings back into a five-point distribution in a way that shows that they understand how other reviewers would bin ratings that were originally from a five-point scale. Given that prior research suggests that people may not create binary categories from a multi-point scale in an objectively neutral way (i.e. by splitting at the midpoint of the scale), could we assign a labelling scheme to the resulting multi-point distribution so that the generated rating distributions account for the asymmetric split point for binary binning? This is the second research question that we aim to answer in this paper.

Pilot Study

Materials & Methods

Participants

A total of 180 participants were recruited in Feb 2023 from Prolific Academic and paid £0.75 each for task completion. Participants were English speakers over 18 years of age. No additional demographic data were collected.

Procedure

As an exploratory investigation into how product evaluations differ when ratings are provided in the form of a binary scale, relative to when they are provided on a five-point scale, we conducted a pilot study. Using a similar format to Ocean (2024), we provided respondents with a series of ten products: a chest of drawers; a fiction novel; wireless earbuds; a smartphone; a movie; a video game; a restaurant, a hotel, a digital camera; an air fryer. These represent an assortment of both search goods and experience goods. Participants were shown an image of the product, followed by ratings information. They were subsequently asked to evaluate product quality on a 0-100 scale and purchase intentions on a seven-point Likert scale ranging from "Extremely unlikely to purchase" to "Extremely likely to purchase".

In the control condition, we presented a typical five-star ratings distribution under each product. This included the mean rating, the total number of ratings, and the percentage of ratings of each score from one to five. The overall mean of the means for each product was 4.04 stars. In the treatment condition, we converted this ratings distribution into a binary "thumbs up" format by splitting the ratings at the midpoint of the scale so that 4-star and 5-star ratings were treated as a "thumbs-up" and 1-star and 2-star ratings were treated as a "thumbs-down". Ratings of 3-stars were divided equally between the two categories. The full experimental script can be found in the OSF repository for this study: [redacted]

Results and Discussion

Linearly mapping the mean for all 10 products to the 0-100 point quality scale in a naïve fashion (i.e. $quality = 25\mu - 25$) suggests that the mean evaluation of product quality across the whole experiment should be approximately 76 if participants were simply mapping the mean rating directly

to evaluate product quality. Mean product quality was 63.4 (n = 98, 95% confidence interval = [61.2, 65.6]) in the control (5-star) condition and 73.1 (n = 82, 95% confidence interval = [71.4, 74.8]) in the treatment (thumbs-up) condition. A two-tailed two sample t-test of independent means yielded $p < 0.00005$ and Cohen's $d = 1.02$. Similarly, for purchase likelihood on a 1 - 7 scale, the means for each condition were 4.24 (95% confidence interval = [4.07, 4.40]) and 4.65 (95% confidence interval = [4.45, 4.85]) respectively, yielding a t-test p-value of $p = 0.002$ and effect size $d = 0.472$. The means and confidence intervals are plotted in Figure 1. The standard errors are likely to be underestimates due to the fact that variances for evaluations of the same product are likely to be correlated even across individuals. However, an OLS regression controlling for product fixed effects and clustering standard errors by product also finds significant treatment effects for both perceived quality ($\hat{\beta} = 9.73$, $p = 0.0004$) and purchase intentions ($\hat{\beta} = 0.411$, $p = 0.0058$). We do not report regression results in detail for the pilot, however we do so later for our main studies.

Overall, the pilot results suggest that there is strong initial evidence to support the fact that products that are rated on an 'equivalent' binary scale are evaluated more highly than when they are rated on a 5-star scale. In other words, this suggests that it is likely that people do not treat the ratings scale linearly. The results imply that for participants to value the products equally in both cases, one would need to reduce the 'bin size' for what constitutes a "thumbs-up" because only the uppermost ratings on a numerical scale are treated as signals of high quality. However, we do not yet know where the 'cut point' lies such that products in the two rating systems are evaluated equivalently.

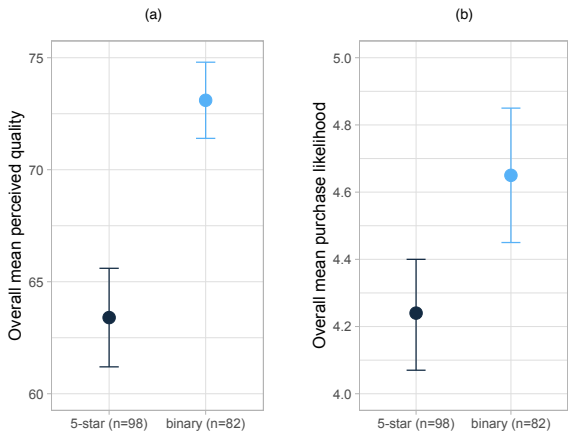

Figure 1: Results from pilot study. Overall means and confidence intervals across all 10 products by treatment. Each panel refers to a different dependent variable: (a) perceived product quality; (b) purchase likelihood.

Study 1

Materials & Methods

Participants

Based on a moderate effect size of $d = 0.5$ with power = 0.95, the minimum cell size for significance at $\alpha = 0.05$ is 105, i.e. a minimum of 420 across four conditions. We recruited 554 participants in total via Prolific in Summer 2024 (319 female, 227 male, 5 non-binary, 3 not reported). Participants were randomly allocated to one of four conditions, and were distributed as follows: *control* = 154, *mid-split* = 135, *low-split* = 126, *high-split* = 139. Mean age in the sample was 35.8 (min = 18, max = 77, $\sigma = 12.48$). Participants were each paid £1.05 for completion of the online experiment. Mean completion time was 3.93 minutes.

Procedure

Study 1 was designed to test the threshold at which individuals categorise ratings from a continuous 5-star scale into binary 'good' and 'bad' bins to determine product quality. The flow of the study was the same as the pilot. Participants were sequentially presented with the following 10 products: Bluetooth speakers; chest of drawers; air fryer; electric kettle; book; badminton set; body lotion; shoes; alarm clock; laptop bag. In contrast to the pilot, we focused on retail products rather than including experiences such as hotels, which may be subject to a slightly different evaluation process. Each product screen contained a brief product name along with an image of the product, and finally the associated ratings information. Participants were asked to rate perceived product quality on a 0-100 scale as in the pilot study and in Ocean (2024). They were also asked to state their likelihood of purchasing the product based on the displayed information using a 7-point Likert scale. The use of these two distinct measures allowed us to capture both people's perception of inherent product quality, as well as the personal preferences that drive purchase intentions. The order of products was randomly selected for each participant. Participants also provided demographic information (age and gender) before completing the main product evaluation task. The full experimental script can be found in the OSF repository for this study: [redacted]

We omitted information on price and branding so that evaluations would be based solely on ratings, and ensured that there was enough variability in average ratings and ratings distributions to cover a variety of contexts. The total volume of ratings was fixed at 200 across all products to avoid interactions between ratings volume and the relative weighting applied to a rating (people are likely

to place higher weight on a given ratings distribution being an accurate signal of product quality when that distribution is being generated by a high volume of total ratings).

Study 1 extended the general design of the pilot by introducing different binary split-points. Participants were randomly assigned to one of four conditions. In the *control* condition, participants were shown review information using the standard five-star format. The mean rating was shown, along with the total volume and the percentage distribution of ratings across the five points of the scale. In the *mid-split* condition, we replicated the treatment from the pilot study to convert the five-point scale ratings into a 'thumbs-up' rating by taking ratings of four and five and half of the three-star ratings as positive. In addition to the mid-split condition from the pilot, we added two further treatments where we changed the split point, i.e. the point at which we draw the line between 'positive' and 'negative' ratings. The *low-split* condition took all ratings of three and above as positive, while the *high-split* condition took only ratings of four and above as positive. These conditions explore the potential for individuals' subjective neutral point to lie either side of the midpoint of the scale.

Results

We first computed the mean reported product quality and mean likelihood of purchase across all 10 products per individual. These means are plotted in Figures 2a and 2b, along with 95% confidence intervals.

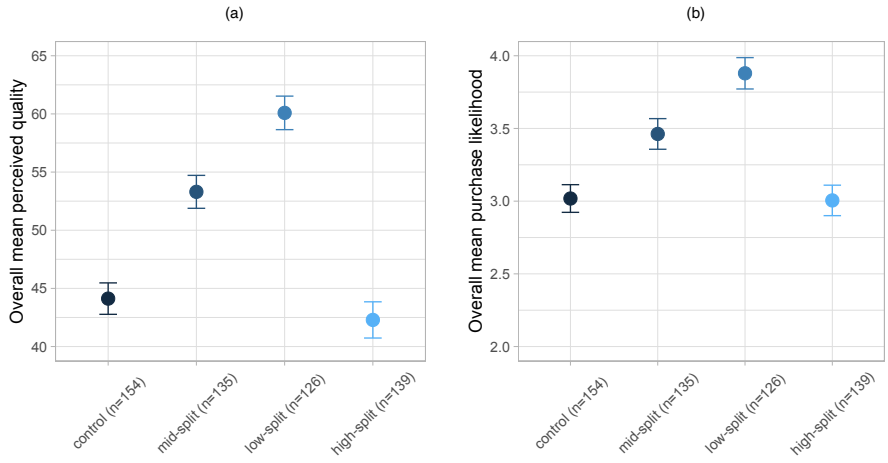

Figure 2: Results from Study 1. (a) Overall mean perceived product quality across all 10 products per participant. (b) Overall mean likelihood of purchase over all 10 products per participant. Bars represent 95% confidence intervals.

As discussed in the analysis for the pilot study, a naïve t-test is likely to contain biased standard errors because of the experiment design involving multiple product evaluations per individual. Therefore, to determine average treatment effects on perceived product quality and on purchase likelihood more formally, we estimated linear regression models. The base specification of the model is:

$$y_{ij} = \beta_0 + \beta_1 treatment_{ij} + \beta_2 product_{ij} + \beta_3 X_{ij} + \epsilon_{ij}$$

for i = 1, 2, …, 554 (individuals) and j = 1, 2, …, 10 (products), where y_{ij} represents the dependent variable (either perceived quality or purchase likelihood) and X_{ij} is a matrix of demographic variables (i.e. age and gender in the present case).

We used two different estimation methods to correct standard errors because of the nested structure of the data (i.e. ten products per individual). First, we used OLS with robust standard errors that were clustered by product (allowing for error terms to be correlated within the same product). Second, we estimated a mixed effects model with a random intercept for each participant. This adds a separate error term u_i to the model specification to account for individual specific effects.[1] Both estimation methods give identical parameter estimates because they are both linear models that are estimating the same specification. However, because the error terms are modelled differently, standard errors will differ across the approaches. The results using both approaches are shown in Table 1.

Table 1 shows that regardless of the estimation method used, and regardless of whether we include additional demographic controls, we see a similar pattern for treatment effects. Perceived quality is deemed to be significantly greater when ratings are displayed in a binary fashion, compared with when they are displayed on a 5-star scale, but *only when the ratings are split at or below the mid-point of the scale*. When ratings are split at the mid-point, perceived quality is approximately 9 points higher. When ratings are split between 2-stars and 3-stars, perceived quality is approximately 16 points higher. However, there is no significant difference between quality using binary and 5-star scales in the *high-split* condition, i.e. when ratings are binned into positive only above 3-stars. This pattern is preserved even when we use purchase likelihood as the dependent variable instead of perceived quality.

[1] We also tested a model that included random intercept and slope parameters, however a likelihood ratio test found that the slope term did not add additional explanatory power.

Table 1: Linear regressions estimating treatment effects in Study 1

	OLS (with s.e. clustered by product)				Mixed model (with random intercept per person)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent variable:	Quality	Quality	Likely	Likely	Quality	Quality	Likely	Likely
Mid-split	9.179***	9.204***	0.444***	0.458***	9.179***	9.204***	0.444***	0.458***
	(1.294)	(1.286)	(0.117)	(0.119)	(1.173)	(1.178)	(0.101)	(0.100)
Low-split	15.97***	15.97***	0.861***	0.868***	15.97***	15.97***	0.861***	0.868***
	(3.040)	(3.050)	(0.228)	(0.229)	(1.195)	(1.199)	(0.103)	(0.102)
High-split	-1.832	-1.805	-0.0131	-0.0045	-1.832	-1.805	-0.0131	-0.0045
	(2.767)	(2.749)	(0.129)	(0.130)	(1.164)	(1.168)	(0.100)	(0.0994)
Includes fixed effects per product	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Includes controls for age and gender	No	Yes	No	Yes	No	Yes	No	Yes
Constant	71.52***	70.78***	4.809***	5.044***	71.52***	70.78***	4.809***	5.044***
	(0.794)	(1.037)	(0.0741)	(0.0931)	(0.979)	(1.589)	(0.0829)	(0.135)
Random intercept variance					79.48	79.17	0.600	0.583
					(5.957)	(5.938)	(0.0440)	(0.0430)
Observations	5540	5540	5540	5540	5540	5540	5540	5540
R-squared	0.657	0.657	0.515	0.52				
Number of groups					554	554	554	554

Notes: (1)-(4) estimate treatment effects with OLS with robust standard errors clustered by product. (5)-(8) estimate treatment effects using a mixed (multilevel) model with a random intercept per individual. *** indicates $p < 0.01$. The dependent variables are: Quality = perceived product quality on 0-100 scale; Likely = likelihood of purchasing product based on ratings information on 7-point Likert scale.

Discussion

Study 1 sought to investigate how consumers mentally categorize product ratings and determine the optimal mapping of continuous rating scales to binary categories. The results clearly and unambiguously suggest that products are evaluated equivalently regardless of whether the ratings are displayed in binary or 5-star formats *only when the binary display uses a cut-point that is above the midpoint of the scale*. This suggests that when consumers see five-point ratings distributions, it is likely that they are categorising 4-star and 5-star ratings as positive, and any rating below 4-stars as negative. This is consistent with Dolnicar and Grün's (2013) finding that individuals regard Likert

responses around the midpoint of a five-point scale as equivalent to negative responses on a binary response scale.

Study 2

Materials & Methods

Participants

We recruited 554 English speakers over 18 years of age via Prolific in Summer 2024 (275 male, 254 female, 5 non-binary, 3 preferred not to say, 1 non-response). Mean age was 30.6 (min = 18, max = 70, σ = 10.98). As in Study 1, we aimed for a minimum group size of 105. Participants were randomly allocated to one of four conditions, and were distributed as follows: *control* = 135, *verbal-balanced* = 135, *verbal-negative* = 136, *verbal-positive* = 132. Participants were each paid £1 for task completion. Mean completion time was 8.42 minutes.

Procedure

Study 2 was designed to answer two questions. First, we wanted to understand how consumers translate a binary ratings profile back into a five-point distribution to establish what kind of rating distribution individuals envisage as forming a particular binary categorisation. Second, we wanted to establish how changing the labelling on a five-point scale to change the perceived neutral point of the scale can moderate this mapping, so that we can find the scale interpretation that is closest to the high-split binning observed in Study 1.

As in Study 1, each participant was shown 10 different product screens. Each screen contained the product name and an image, along with rating information that showed the percentage of positive and negative reviews. Participants were told that the products were originally rated on a five-point scale but this had been converted to a binary format for presentation purposes. The task was for participants to try to reconstruct the original ratings distribution by allocating 100 ratings to each of the five points of the scale. They did this by moving five independent sliders that were constrained to a sum of 100. For convenience, the original distributions we converted were the same as the 10 distributions constructed for the control condition of Study 1, though we attached these distributions to 10 new products. The full list of products and experiment script can be found in the OSF repository

for this study: [redacted]. The distributions were converted using the neutral *mid-split* method from Study 1 (participants were not given any information on how the ratings were converted). The resulting positive to negative rating ratios for each of the 10 products using this method were: (1) 85:15, (2) 72:28, (3) 58:42, (4) 44:56, (5) 28:72, (6) 76:24, (7) 97:3, (8) 18:82, (9) 78:22, (10) 43:57. With a midpoint split, we might expect that people's reconstruction of the five-point scale would be skewed more towards higher ratings than the source distribution. This is because individuals may not consider that some ratings at the midpoint could spill over into either of the binary bins if there is a two-way equivalence between ratings of only four and five being interpreted as positive. If this is true, then the mean of the reconstructed distribution should therefore also be greater than the source distribution. On the other hand, if the mean of the reconstructed distribution is not significantly different from the mean of the source distribution, then it is possible that individuals exhibit a skew when categorising multiple points into binary bins, but do not account for this when going in the other direction, i.e. when expanding binary categories to a multi-point scale.

To test the second question, we randomly allocated participants to four groups. The control condition asked participants to expand the binary ratings to a typical unlabelled five-star rating scale. We then designed three treatment conditions to determine whether verbal labelling of scales can moderate the reconstructed distribution mapping. The *verbal-balanced* condition replaced the five points with the following set of five labels (from best rating to worst rating): {*amazing, very good, average, poor, terrible*}. These labels were chosen as they correspond to those used on TripAdvisor.com, and emphasise that the neutrality of the midpoint. Specifically, the top two scale points are assigned positively-valenced labels, the middle scale point is assigned a neutral label, and the bottom two scale points are assigned negatively-valenced labels. This treatment was designed to be a baseline verbal conversion from the 5-star scale, and so we did not expect any difference in reconstructed distributions between this condition and the control condition. The *verbal-negative* treatment replaced the five points with the following set of five labels (from best rating to worst rating): {*very good, average, poor, terrible, atrocious*}. The *verbal-negative* treatment takes the *verbal-balanced* treatment and shifts the neutral label up to the fourth point of the scale. In effect, this corresponds more closely to how individuals appeared to categorise a five-point distribution into binary bins in Study 1. Finally the *verbal-positive* treatment replaced the five points with the following set of five labels (from best rating to worst rating): {*incredible, amazing, very good, average, poor*}. The *verbal-positive* treatment shifts the neutral label down to the second point of the scale.

These treatments are somewhat analogous to the treatments in Study 1 where we manipulated the split point in the conversion of 5-star to binary ratings. The results from Study 1 suggest that people

are binning ratings into positive and negative categories by splitting *above* the mid-point of the scale. This means that while providing a balanced scale is likely to lead to a reconstructed distribution that is more positively-valenced than the original (i.e. with a higher mean), providing a shifted scale may lead to a less biased reconstruction of the source distribution. In particular, by shifting the neutral point down as in the *verbal-positive* treatment, positive ratings may be more evenly distributed across the range, thereby lowering the reconstructed mean score and bringing it closer to the original mean rating than what we would expect from the other three treatment conditions.

Results

Figure 3 plots the average estimated distributions for each product under each treatment condition alongside the source distribution from which the binary rating was obtained. Overall, we see that participants tend to translate binary scores into unimodal distributions, regardless of the labelling used for the rating scale. This means that participants did not seem to recognise the tendency for people to either 'brag' or 'moan' and therefore did not produce a bimodal density function concentrated at the extremes. Although they were reasonably accurate in approximating the spread of opinions when the positive to negative ratio was relatively disparate, they were not so accurate in judging distributions when the ratio was similar (i.e. for products with a mixed opinion, such as Product 3). Figure 3 suggests that people are likely to interpret a relatively even ratio of positive to negative ratings as a mostly uniform spread across the range rather than as a bipolar distribution. Participants also did not recognise the prevalence of the J-shape in ratings distributions but did tend to generate skewed distributions.

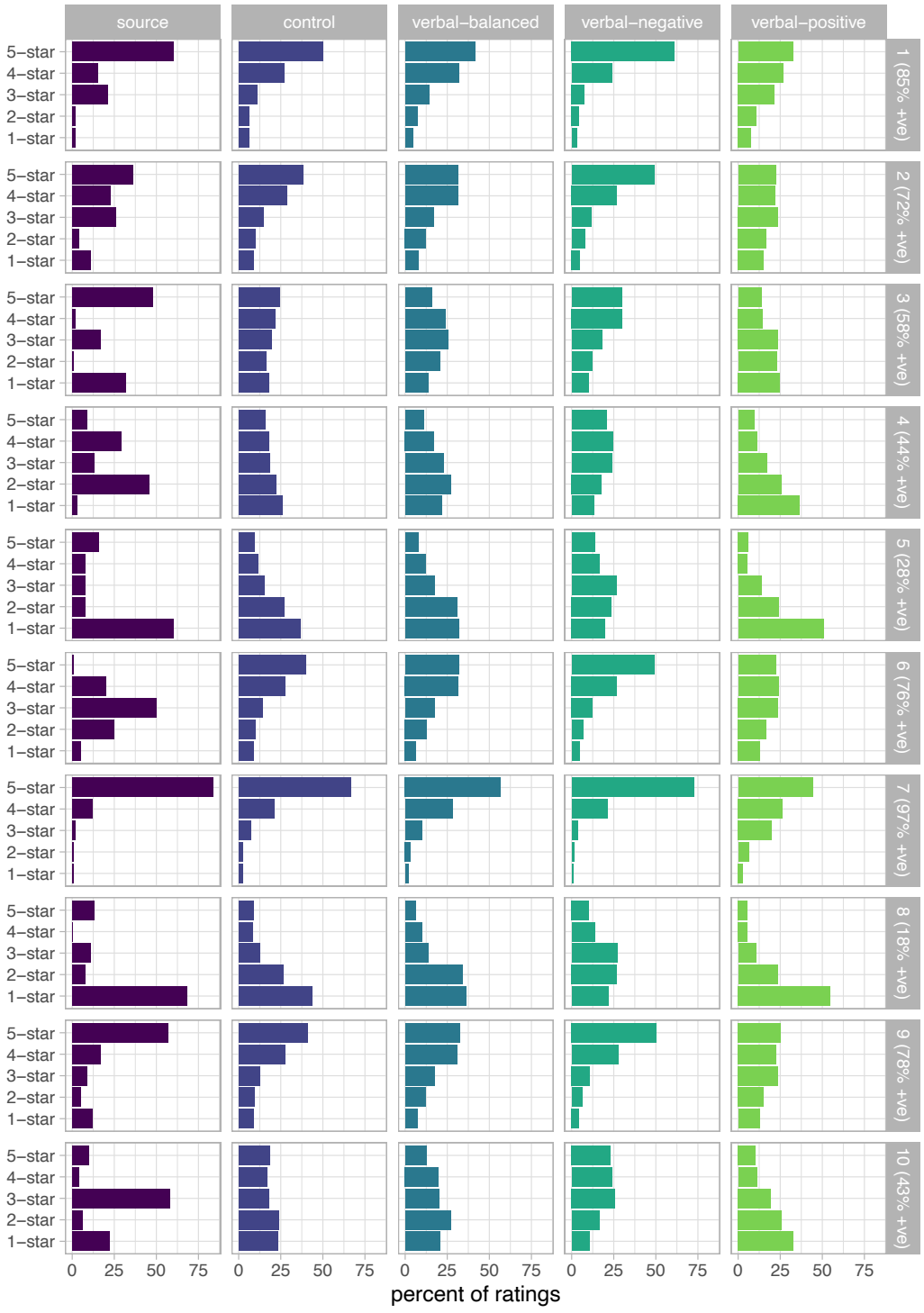

Figure 3: Participant estimated source distributions from binary ratings, by treatment (columns) and by product (rows)

Table 2 shows the first three moments of the source and estimated distributions for each product as well as for the averaged distribution across all products. The *verbal-positive* and *verbal-negative* treatments skew the estimated distributions further from the skew of the averaged source distribution. We can also see this visually in Figure 4, which plots the densities of the averaged distributions over all products for each treatment. While the estimated mean of the distribution is slightly higher in the control condition than the source distribution, the mean for the *verbal-balanced* condition is slightly lower overall. Therefore, we do not observe a consistently upwards biased mean when binary ratings are converted to a 5-point scale, apart from in the *verbal-negative* condition. In this treatment, only Product 7 had a lower estimated mean than the original source distribution, and this is likely because of noise at the extremes.

Table 2: Moments of estimated distributions relative to source for each treatment

	Mean					Standard Deviation					Skewness				
Product	source	con	v-b	v-n	v-p	source	con	v-b	v-n	v-p	source	con	v-b	v-n	v-p
1 (85:15)	4.29	4.09	3.99	4.35	3.66	0.99	1.17	1.13	1.01	1.25	-1.22	-1.29	-1.06	-1.74	-0.62
2 (72:28)	3.69	3.77	3.66	4.07	3.19	1.29	1.29	1.26	1.16	1.36	-0.74	-0.83	-0.68	-1.19	-0.19
3 (58:42)	3.33	3.18	3.07	3.58	2.70	1.77	1.43	1.28	1.30	1.35	-0.33	-0.18	-0.07	-0.61	0.29
4 (44:56)	2.95	2.74	2.68	3.22	2.32	1.11	1.42	1.29	1.31	1.32	0.36	0.23	0.30	-0.21	0.69
5 (28:72)	2.12	2.30	2.33	2.81	1.92	1.56	1.32	1.25	1.30	1.18	0.94	0.72	0.67	0.20	1.23
6 (76:24)	2.87	3.79	3.69	4.09	3.26	0.82	1.30	1.22	1.14	1.33	-0.20	-0.86	-0.67	-1.21	-0.24
7 (97:3)	4.77	4.47	4.35	4.64	4.03	0.63	0.92	0.92	0.71	1.08	-3.61	-2.02	-1.57	-2.48	-0.92
8 (18:82)	1.82	2.13	2.16	2.64	1.84	1.39	1.30	1.20	1.25	1.16	1.49	0.97	0.91	0.35	1.39
9 (78:22)	4.02	3.82	3.69	4.13	3.33	1.39	1.30	1.24	1.12	1.34	-1.21	-0.90	-0.69	-1.31	-0.30
10 (43:57)	2.74	2.84	2.77	3.33	2.40	1.15	1.43	1.32	1.29	1.33	0.04	0.19	0.22	-0.27	0.61
All	3.26	3.31	3.24	3.69	2.86	1.53	1.50	1.40	1.34	1.46	-0.26	-0.31	-0.22	-0.66	0.12

Note: con = control, v-b = verbal-balanced, v-n = verbal-neutral, v-p = verbal-positive. Ratios in parentheses represent the percentages of 'thumbs-up' relative to 'thumbs-down shown to participants for each product.

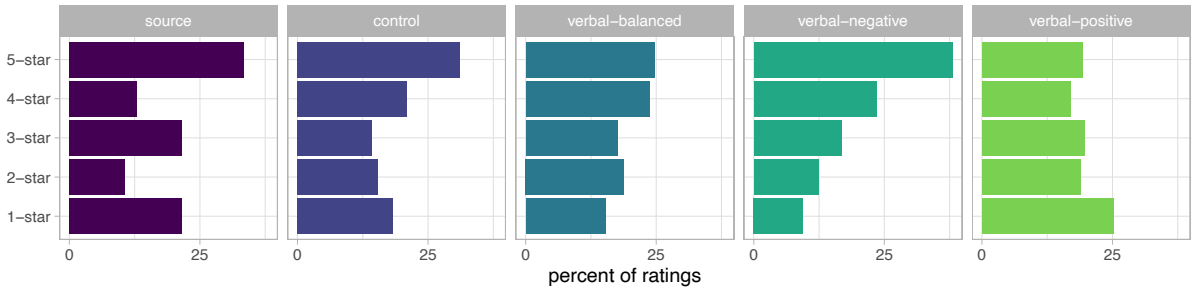

Figure 4: The overall average of the reconstructed distributions by treatment condition, relative to the average source distribution

To systematically analyse the differences between source and estimated distributions, we conducted Kolmogorov-Smirnov (K-S) tests. Table A1 presents the K-S test statistics for each product, under the null hypothesis that the source distribution is not significantly different from the generated distribution. Higher test statistics (lower *p*-values) indicate greater divergence from the original source distributions. The results suggest that on average, participants generated distributions that were closer to the source when asked to do so on the regular five-point scale, compared with any of the verbally labelled scales. Figures A1 and A2 plot the source and estimated cumulative distributions for each product and treatment, as well as for each treatment overall.

Finally, we compared the proportion of positive ratings participants were shown relative to their estimated distributions to see how people implicitly categorise ratings into positive and negative, and subsequently translate this into a distribution. Table 3 shows the proportion of 'thumbs-up' ratings shown for each product, and the mean percentage of positive ratings derived from the estimated distributions. We use three different definitions of positive, based on the three different split points used in Study 1. The high-split takes only 4-star and 5-star ratings as positive, the mid-split also includes half of the 3-star ratings, and the low-split takes all 3-star ratings and above as positive. The results show that in the neutrally-valenced conditions (control, *verbal-balanced*) the estimated distributions follow the mid-split definition of a positive rating most closely. In the *verbal-negative* treatment, estimated distributions follow the high-split definition of a positive rating most closely. In the *verbal-positive* treatment, estimated distributions follow the low-split definition of a positive rating most closely. The average 'thumbs-up' to 'thumbs-down' ratio across all 10 products of 60:40 corresponds most closely to the average ratio observed in the distributions generated under the *verbal-negative* treatment of 62:38.

Table 3: Mean proportion of 'positive' ratings in Study 2 estimated distributions based on the cut points from Study 1

Product	Split point	control (% +ve)	verbal-balanced (% +ve)	verbal-negative (% +ve)	verbal-positive (% +ve)
1 (85% +ve)	High-split	77.24	73.67	84.93*	59.59
	Mid-split	82.62*	80.87	88.76	70.41
	Low-split	88.01	88.07*	92.58	81.23*
2 (72% +ve)	High-split	66.46	62.81	75.88*	44.32
	Mid-split	73.84*	71.24*	81.61	56.17
	Low-split	81.21	79.68	87.34	68.03*
3 (58% +ve)	High-split	45.98	39.84	59.89*	28.45
	Mid-split	55.70*	52.56*	68.92	40.37

	Low-split	65.42	65.28	77.94	52.30*
4 (44% +ve)	High-split	33.39	28.07	45.19*	20.78
	Mid-split	42.51*	39.47*	57.20	29.34
	Low-split	51.62	50.86	69.21	37.91*
5 (28% +ve)	High-split	21.03	19.76	30.09*	11.29
	Mid-split	28.66*	28.57*	43.33	18.26
	Low-split	36.29	37.38	56.58	25.23*
6 (76% +ve)	High-split	67.33	63.40	75.96*	46.48
	Mid-split	74.34*	72.03*	82.19	58.30
	Low-split	81.34	80.66	88.42	70.11*
7 (97% +ve)	High-split	87.70	84.84	94.09	70.62
	Mid-split	91.41	89.82	95.82	80.64
	Low-split	95.13*	94.80*	97.56*	90.65*
8 (18% +ve)	High-split	17.31*	16.24*	24.09*	11.07
	Mid-split	23.66	23.02	37.66	16.39*
	Low-split	30.01	29.80	51.23	21.70
9 (78% +ve)	High-split	68.76	63.10	78.27*	48.25
	Mid-split	75.12*	71.91	83.70	60.06
	Low-split	81.49	80.71*	89.13	71.86*
10 (43% +ve)	High-split	35.49	32.36	47.15*	21.70
	Mid-split	44.32*	42.44*	60.08	31.39
	Low-split	53.16	52.53	73.00	41.07*
All (60% +ve)	High-split	52.07	48.41	61.55*	36.26
	Mid-split	59.22*	57.19*	69.93	46.13
	Low-split	66.37	65.98	78.30	56.01*
n		135	135	136	132

Note: The numbers are the percentage of 'positive' ratings rounded to 2.d.p. For high-split, positive ratings are defined as either 4-star or 5-star. For mid-split, positive ratings are defined as 4-star, 5-star, and half of all 3-star ratings. For low-split, positive ratings are defined as 3-star, 4-star, or 5-star. Percentages in parentheses next to product numbers represent the percentages of 'thumbs-up' reviews as shown to participants in the task. * represents the split that most closely matches the ratio that participants were given.

Discussion

Study 2 finds that participants appear to be poor at recreating the source distribution when shown only a binary version of it, apart from in cases where it is unimodal and skewed towards higher ratings. Their tendency to generate smooth distributions suggests that individuals tend to underestimate the degree of noise / variation in other people's opinions. Attaching verbal labels to points of the scale instead of numerical values did little to change estimated distributions when the valence of the words was kept neutral. However, verbally skewing the labels appeared to worsen rather than improve fit. This suggests that recalibrating the perceived midpoint of the scale to account for the asymmetry in binary bias observed in Study 1 does not seem to improve correspondence between estimated and source distributions (although it may do in specific cases).

Second, if our finding from Study 1 were also to hold in reverse, then when asked to map binary ratings to a five-point distribution, we would expect individuals to assign positive reviews a rating of either four or five and negative reviews a rating of one, two, or three. The results from Study 2 showed that this happens only in the *verbal-negative* treatment, i.e. when the neutral point scale was deliberately shifted so that the fourth point on the scale was labelled as 'average'. This generates an interesting and slightly paradoxical conclusion. When people see a distribution of ratings on a five-point scale, they consider only ratings of four or five to be positive signals of quality. However, when people see ratings on a binary scale, then they do not appear to infer that binning would occur anywhere other than the midpoint *unless* we explicitly define a non-midpoint split point by labelling the scale so that the neutral point lies away from the midpoint.

General Discussion

The present studies set out to investigate the phenomenon that consumers appear to value products differently based on whether ratings are provided to them on a five-point scale or on a binary (i.e. positive and negative) scale. Overall, there are two main findings. First, there is a consistent tendency to value a product more highly when ratings are presented on a binary scale as opposed to a five-point scale when the mapping from five-point to binary is exactly at the midpoint (i.e. when ratings of one and two are considered negative, ratings of four and five are considered positive, and ratings of three are allocated evenly between positive and negative). Furthermore, we find that the binary mapping that results in equivalent product valuations is *above* the midpoint. In

other words, when ratings of only 4 and 5 are considered as 'positive', products will be valued equivalently between the two scales. This implies that *consumers view a rating at the midpoint of a scale as negative rather than neutral*. This finding remained consistent after accounting for individual differences and demographic factors.

Second, our findings suggest that people do not anticipate or account for this 'above the midpoint' bias when presented with binary ratings to begin with. Instead, when consumers are provided with only binary ratings, they appear to implicitly assume that the underlying distribution considers a positive review to be anything at the midpoint of the scale or above. Only when we shift the valence of the five-point scale downwards by labelling a rating of four as 'average', therefore explicitly defining four as a neutral point, are individuals correctly able to adjust for this bias and generate a distribution that more closely represents how others actually categorise positive and negative ratings.

These findings contribute to the literature on binary bias (e.g. Fisher et al., 2018) as well as to the wider understanding of categorical thinking in consumer behaviour that illustrates how simplification strategies might influence product evaluations and purchase intentions (e.g. Gutman, 1982; Mogilner et al., 2008). Prior work by Ocean (2024) found evidence that the weights placed on ratings of three, four, and five on a five-point scale implied that they sent a broadly equivalent positive signal of product quality to consumers. However, the fact that individuals who see only binary reviews that are generated by splitting the five-point scale at the midpoint to generate 'positive' and 'negative' bins leads individuals to perceive *higher* product quality than when viewing the entire distribution suggests that this symmetric binning overestimates the number of positive signals of product quality that the full distribution is sending. This asymmetry may be a form of negativity bias (Rozin & Royzman, 2001) because it suggests that consumers require more strongly-valenced positive signals to offset any indications of low product quality.

It appears that industry has already implicitly assumed this phenomenon to be true. For example, Reichheld's Net Promoter Score for company recommendations only considers scores of 9 or 10 on a 10-point scale to be a positive recommendation, score of 7 or 8 to be neutral, and any score below 7 to be a negative recommendation (Reichheld, 2011). Businesses and customers already seem to be attuned to the idea that truly positive ratings require surpassing a threshold higher than the numerical midpoint of a scale (Koh et al., 2010). Schoenmueller et al. (2020) conducted a comprehensive examination of online product ratings and found evidence of "ratings bubbles" – the fact that average product ratings have consistently increased over time across several platforms, leading to a compression of ratings at the top of the scale. Whether or not consumers realise that the reason for this is likely to be because of a self-selection bias in reviewers (Schoenmueller et al., 2020),

i.e. the tendency to either 'moan' or 'brag' about products (e.g. Hu et al., 2006), it appears that they have come to expect high ratings as the norm.

There is an asymmetry implied by our results, which suggests different courses of action depending on the goal. Retailer websites that wish to summarise ratings information and also benefit from product sales will have an incentive to bin reviews into positive and negative by splitting exactly at the midpoint, because this will increase product valuation. This also highlights possible areas for regulatory involvement, as firms may use this fact to mislead consumers into making undesirable purchases. From a policy perspective, improving consumer welfare in online shopping may involve requiring firms to provide clear visual or verbal indicators to highlight the difference between ratings above and below the perceived inflection point, or perhaps mandating the provision of different rating summaries at different stages of the online shopping process (Chen, 2017). On the other hand, review aggregator websites that benefit from user engagement rather than product sales have more of an incentive to maintain unbiasedness. Therefore, they should choose to emphasise the inherent skew in the interpretation of the rating scale, either by binning above the midpoint if converting to binary, or labelling rating scales to emphasise a neutral point that is above the numerical midpoint of the scale. Finally, consumers may benefit just by being aware of this bias in others' ratings because they will be better able to infer the true perceived valuation of a product regardless of which rating scale is used.

The main limitation with the present studies is that there was no incentive compatibility or actual purchase behaviour inherent within the design. Therefore, although we think that our results provide valuable insight into consumer psychology and judgements, actual behaviour may differ in field settings based on context, especially when high value purchases are being made. Future research would benefit from website data on purchases or click-through rates to more directly analyse the impact of rating presentation on consumer activity. Our studies used a broad set of products categories that spanned commonly purchase product types on online shopping websites, but did not specifically seek to test differences in effect across different product categories. Future studies may seek to assess whether the effects we observed hold across a broader range of product types, particularly for experience goods versus search goods, as the influence of ratings may differ between these categories (e.g. Mudambi & Schuff, 2010). Furthermore, our experimental design presented ratings in isolation. In real-world e-commerce settings, ratings are frequently complemented by written reviews. Future research could look into whether the interaction between ratings and written reviews changes how ratings are translated between binary and continuous scales. The studies did not investigate cultural background or familiarity with online purchasing. While our aim was to

understand consumer behaviour more generally, we accept that there may be differences in perception of scales and rating behaviour across cultures (Koh et al., 2010).

Conclusion

Overall, this paper finds that when a five-point rating scale is translated into a binary scale by binning ratings into two groups (i.e. positive reviews and negative reviews) with a neutral point at the midpoint of the scale, individuals will perceive the product as having higher quality when they assess the product from the binary classification than when they assess the product from the five-point distribution. To translate the five-point scale to a binary scale without affecting product evaluations, one must classify only ratings of four and five as positive. Furthermore, people do not seem to be aware of this asymmetry if they are asked to estimate a five-point rating distribution when shown ratings on a binary scale. Instead, they generate distributions that imply they are taking the midpoint of the scale as the cut point between positive and negative ratings. Individuals will only be able to identify an asymmetric cut point if the five-point scale they are asked to generate assigns verbal labels to each point of the scale that suggest a rating of four, not three, is neutral.

These findings highlight an important behavioural bias in the interpretation of online reviews that potentially has significant implications for both consumer welfare as well as firm behaviour. The findings are of particular relevance to policymakers and regulators seeking to understand consumer biases to minimise consumer exploitation by firms. Future research should seek to verify these findings in secondary data from e-commerce websites that actually captures real-world decisions in the field, as well as understanding how well our result generalises across product categories, different types of consumers, as well as across cultures. Finally, psychological research should seek to determine whether this phenomenon reproduces when the number of points in the multi-point scale is increased.

References

Barberis, N., & Shleifer, A. (2003). Style investing. *Journal of Financial Economics*, *68*(2), 161–199. https://doi.org/10.1016/S0304-405X(03)00064-3

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Chen, C.-W. (2017). Five-star or thumbs-up? The influence of rating system types on users' perceptions of information quality, cognitive effort, enjoyment and continuance intention. *Internet Research*, *27*(3), 478–494. https://doi.org/10.1108/IntR-08-2016-0243

Chevalier, J. A., & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, *43*(3), 345–354. 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

Dehaene, S. (2011). *The Number Sense: How the Mind Creates Mathematics, Revised and Updated Edition*. Oxford University Press, USA.

Dolnicar, S., & Grün, B. (2013). "Translating" between survey answer formats. *Journal of Business Research*, *66*(9), 1298–1306. https://doi.org/10.1016/j.jbusres.2012.02.029

Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *BMJ*, *324*(7339), 729–732. https://doi.org/10.1136/bmj.324.7339.729

Etumnu, C. E., Foster, K., Widmar, N. O., Lusk, J. L., & Ortega, D. L. (2020). Does the distribution of ratings affect online grocery sales? Evidence from Amazon. *Agribusiness*, *36*(4), 501–521. https://doi.org/10.1002/agr.21653

Fisher, M., & Keil, F. C. (2018). The Binary Bias: A Systematic Distortion in the Integration of Information. *Psychological Science*, *29*(11), 1846–1858. https://doi.org/10.1177/0956797618792256

Fisher, M., Newman, G. E., & Dhar, R. (2018). Seeing Stars: How the Binary Bias Distorts the Interpretation of Customer Ratings. *Journal of Consumer Research*. https://doi.org/10.1093/jcr/ucy017

Gavilan, D., Avello, M., & Martinez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. *Tourism Management*, *66*, 53–61. https://doi.org/10.1016/j.tourman.2017.10.018

Gu, S., Ślusarczyk, B., Hajizada, S., Kovalyova, I., & Sakhbieva, A. (2021). Impact of the COVID-19 Pandemic on Online Consumer Purchasing Behavior. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(6), 2263–2281. https://doi.org/10.3390/jtaer16060125

Gutman, J. (1982). A Means-End Chain Model Based on Consumer Categorization Processes. *Journal of Marketing*, *46*(2), 60–72. https://doi.org/10.1177/002224298204600207

Harvey, C. (2016). Binary Choice vs Ratings Scales: A behavioural science perspective. *International Journal of Market Research*, *58*(5), 647–648. https://doi.org/10.2501/IJMR-2016-041

Hjeij, M., & Vilks, A. (2023). A brief history of heuristics: How did research on heuristics evolve? *Humanities and Social Sciences Communications*, *10*(1), 64. https://doi.org/10.1057/s41599-023-01542-z

Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of Online word-of-mouth communication. *Proceedings of the 7th ACM Conference on Electronic Commerce*, 324–330. https://doi.org/10.1145/1134707.1134743

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*, *52*(10), 144–147. https://doi.org/10.1145/1562764.1562800

Klockars, A. J., & Yamagishi, M. (1988). The Influence of Labels and Positions in Rating Scales. *Journal of Educational Measurement*, *25*(2), 85–96. https://doi.org/10.1111/j.1745-3984.1988.tb00294.x

Koh, N. S., Hu, N., & Clemons, E. K. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, *9*(5), 374–385. https://doi.org/10.1016/j.elerap.2010.04.001

Lafky, J. (2014). Why do people rate? Theory and evidence on online ratings. *Games and Economic Behavior*, *87*, 554–570. https://doi.org/10.1016/j.geb.2014.02.008

Langan, R., Besharat, A., & Varki, S. (2017). The effect of review valence and variance on product evaluations: An examination of intrinsic and extrinsic cues. *International Journal of Research in Marketing*, *34*(2), 414–429. https://doi.org/10.1016/j.ijresmar.2016.10.004

Lee, S., Lee, S., & Baek, H. (2021). Does the dispersion of online review ratings affect review helpfulness? *Computers in Human Behavior*, *117*, 106670. https://doi.org/10.1016/j.chb.2020.106670

Luca, M. (2016). *Reviews, Reputation, and Revenue: The Case of Yelp.Com* (SSRN Scholarly Paper No. 1928601). Social Science Research Network. https://doi.org/10.2139/ssrn.1928601

Mogilner, C., Rudnick, T., & Iyengar, S. S. (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), 202–215. https://doi.org/10.1086/588698

Mudambi & Schuff. (2010). Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185–200. https://doi.org/10.2307/20721420

Ocean, N. (2024). Weighting ratings: Are people adjusting for bias in extreme reviews? *Journal of Experimental Psychology: Applied*, *30*(2), 391–409. https://doi.org/10.1037/xap0000497

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, *72*(6), 407–418. https://doi.org/10.1037/h0022602

Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. *Annals of Tourism Research*, *50*, 67–83. https://doi.org/10.1016/j.annals.2014.10.007

Reichheld, F. (2011). *The Ultimate Question 2.0 (Revised and Expanded Edition): How Net Promoter Companies Thrive in a Customer-Driven World*. Harvard Business Review Press.

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Schoenmueller, V., Netzer, O., & Stahl, F. (2020). The Polarity of Online Reviews: Prevalence, Drivers and Implications. *Journal of Marketing Research*, *57*(5), 853–877. https://doi.org/10.1177/0022243720941832

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating Scales Numeric Values May Change the Meaning of Scale Labels. *Public Opinion Quarterly*, *55*(4), 570–582. https://doi.org/10.1086/269282

Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management*, *32*(6), 1310–1323. https://doi.org/10.1016/j.tourman.2010.12.011

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. https://doi.org/10.1016/J.COGPSYCH.2005.10.003

Sun, M. (2012). How Does the Variance of Product Ratings Matter? *Management Science*, *58*(4), 696–707. https://doi.org/10.1287/mnsc.1110.1458

Tsekouras, D. (2017). The Effect of Rating Scale Design on Extreme Response Tendency in Consumer Product Ratings. *International Journal of Electronic Commerce*, *21*(2), 270–296. https://doi.org/10.1080/10864415.2016.1234290

Wildt, A. R., & Mazis, M. B. (1978). Determinants of Scale Response: Label versus Position. *Journal of Marketing Research*, *15*(2), 261–267. https://doi.org/10.1177/002224377801500209

Appendix

Table A1: Kolmogorov-Smirnov test statistics comparing estimated distributions in each condition to original source ratings distribution

	control		verbal_balanced		verbal_negative		verbal_positive	
Product	D	p	D	p	D	p	D	p
1 (85:15)	0.100	0.699	0.180	0.078	0.100	0.699	0.273	0.001**
2 (72:28)	0.070	0.967	0.050	1.000	0.162	0.141	0.170	0.111
3 (58:42)	0.238	0.002**	0.320	< 0.0005***	0.220	0.016*	0.340	< 0.0005***
4 (44:56)	0.230	0.010*	0.190	0.054	0.183	0.069	0.330	< 0.0005***
5 (28:72)	0.230	0.010*	0.283	0.001**	0.400	< 0.0005***	0.130	0.367
6 (76:24)	0.465	< 0.0005***	0.432	< 0.0005***	0.552	< 0.0005***	0.252	0.003**
7 (97:3)	0.163	0.008**	0.270	0.001**	0.117	0.495	0.396	< 0.0005***
8 (18:82)	0.240	0.006**	0.320	< 0.0005***	0.460	< 0.0005***	0.135	0.315
9 (78:22)	0.164	0.134	0.250	0.004**	0.080	0.403	0.313	< 0.0005***
10 (43:57)	0.216	0.018*	0.190	0.054	0.330	< 0.0005***	0.316	< 0.0005***
All	0.240	0.006**	0.300	< 0.0005***	0.510	< 0.0005***	0.640	< 0.0005***

Note: The null hypothesis is that there is no difference between the source distribution and the estimated distribution. D is the Kolmogorov-Smirnov test statistic. *** indicates $p < .001$, ** indicates $p < .01$, * indicates $p < .05$. Ratios in parentheses represent the percentages of 'thumbs-up' relative to 'thumbs-down shown to participants for each product.

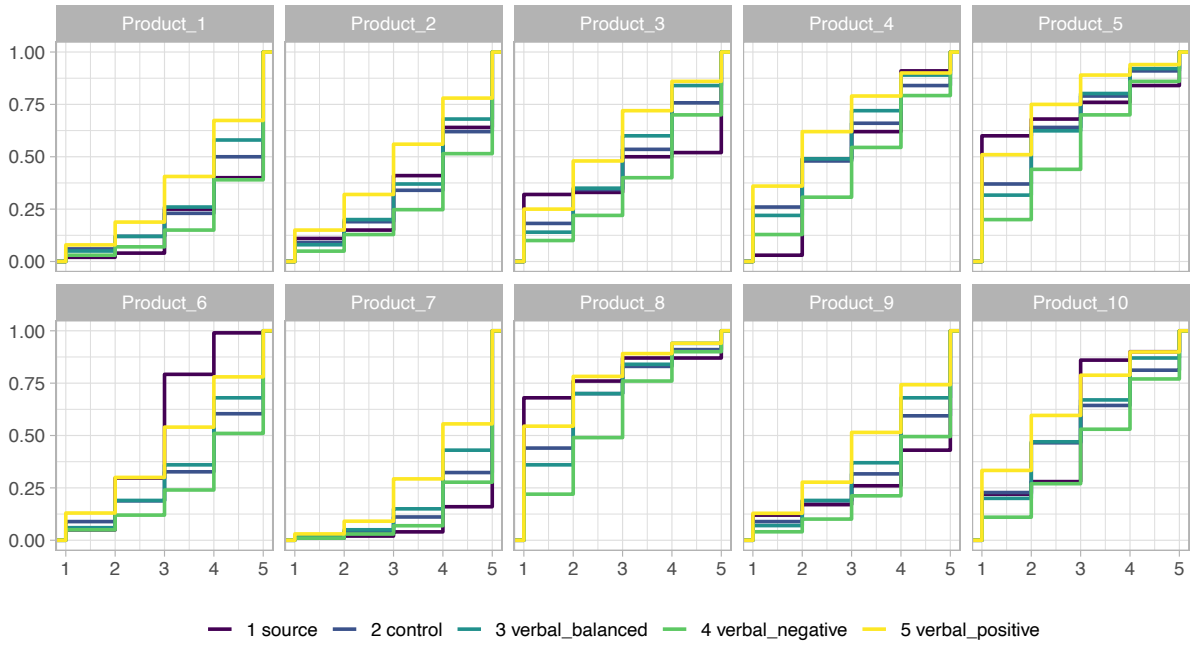

Figure A1: Cumulative distribution functions for source and estimated ratings distributions, by product and treatment

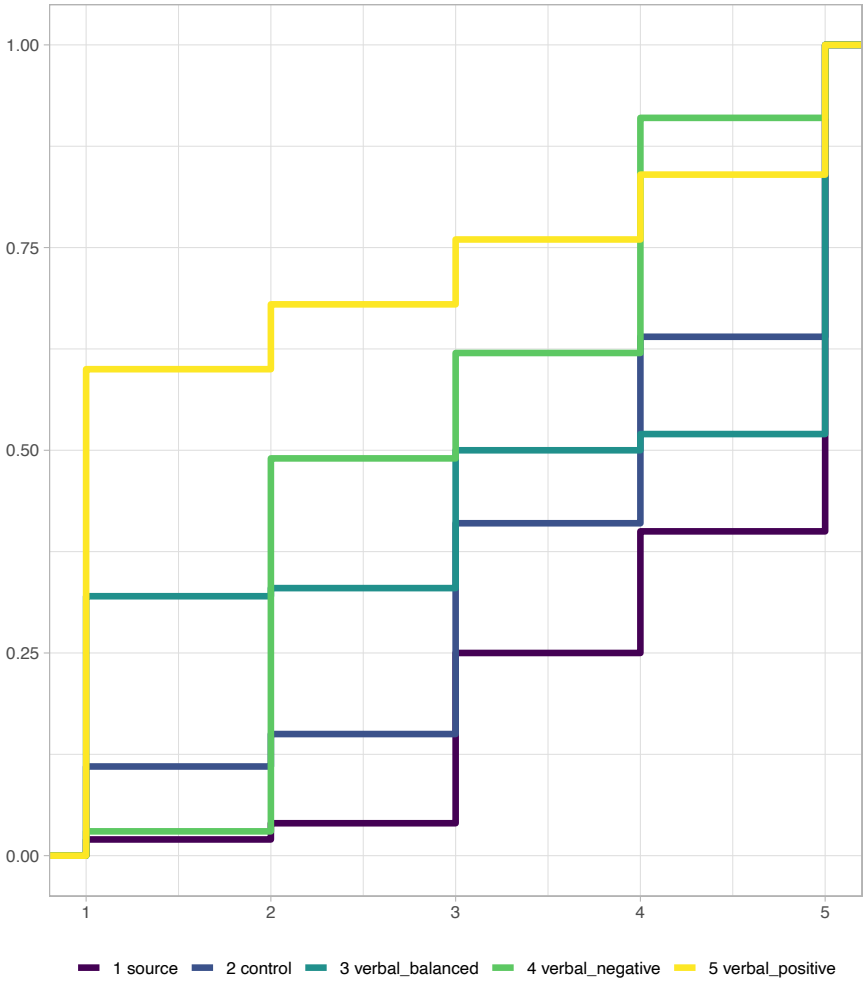

Figure A2: Cumulative distribution functions per condition, averaged across all products